UNIVERSIDADE FEDERAL DO PARANÁ

FLÁVIA DE FÁTIMA COSTA

IDENTIFICAÇÃO DE HOMÓLOGOS E ANÁLISE DE VIZINHANÇA DOS GENES *draTG* ENVOLVIDOS NA REGULAÇÃO DA FIXAÇÃO BIOLÓGICA DE NITROGÊNIO

CURITIBA

2015

FLÁVIA DE FÁTIMA COSTA

IDENTIFICAÇÃO DE HOMÓLOGOS E ANÁLISE DE VIZINHANÇA DOS GENES *draTG* ENVOLVIDOS NA REGULAÇÃO DA FIXAÇÃO BIOLÓGICA DE NITROGÊNIO

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Bioinformática, no Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná.

Orientador: Prof. Dr. Luciano Fernandes Huergo
Coorientador: Prof. Dr. Leonardo Magalhães Cruz

CURITIBA

2015

TERMO DE APROVAÇÃO

FLÁVIA DE FÁTIMA COSTA

"Identificação de homólogos e análise de vizinhança dos genes *draTG* envolvidos na regulação da Fixação Biológica de Nitrogênio"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof . Dr. Luciano Fernandes Huergo

Coorientador: Prof. Dr. Leonardo Magalhães Cruz

Profª Drª Vivian Rotuno Moure
Universidade Federal do Paraná

Profª. Drª. Maria Berenice Reynaud Steffens
Universidade Federal do Paraná

Curitiba, 12 de março de 2015

Dedico este trabalho a minha família
e ao meu marido.

AGRADECIMENTOS

Agradeço a Deus por esta oportunidade.

Ao homem que durante o decorrer do mestrado foi namorado e noivo, e agora se tornou meu marido. Muito obrigada Lucas por toda ajuda, tempo e paciência que dedicou a mim e para a realização deste projeto.

A minha família: meu pai, minha mãe e irmãos pelo amor e carinho recebidos durante toda a vida. A minha nova família: meu sogro, minha sogra e cunhados pela inclusão familiar.

Ao meu orientador Prof. Dr. Luciano Fernandes Huergo pelo acolhimento, acompanhamento e incentivo de todo o trabalho.

Ao meu coorientador Prof. Dr. Leonardo Magalhães Cruz pelas discussões e correções.

Aos professores do Programa de Pós Graduação em Bioinformática pelos ensinamentos.

Aos amigos que fiz: Ana Bandeira, Eduardo Langowski, Venicio Antunes e Calebe Brim, agradeço pelas infinitas e mais variadas conversas que tivemos e também aos demais colegas da Bioinfo.

Aos amigos Ana Paula Sandoval e Rodrigo Sevinhago que abriram a minha mente sobre o mestrado e aos amigos de vida: Bianca, Tião e prima Jaqueline.

Ao órgão financiador da bolsa de estudo CAPES.

E a todos que contribuíram para este trabalho.

Obrigada!

RESUMO

Nitrogenase é a enzima responsável pelo processo de Fixação Biológica de Nitrogênio (FBN), no qual o gás dinitrogênio é reduzido para amônia por microorganismos. Em algumas bactérias fixadoras de nitrogênio, como *Azospirillum brasilense*, a atividade da enzima é regulada ao nível pós-traducional através da ADP-ribosilação. A enzima dinitrogenase redutase ADP-ribosil trasnferase (DraT) catalisa a ADP-ribosilação e, assim, inativa a nitrogenase. Por outro lado, a dinitrogenase redutase glicohidrolase (DraG), remove o grupo modificador e deste modo reativa a nitrogenase. Neste trabalho foram identificadas proteínas semelhantes às enzimas DraT e DraG, depositadas no banco de dados GenBank utilizando softwares de Bioinformática, em especial utilizando BlastP e CD-HIT. As análises revelaram que: os homólogos ao gene *draT* estão restritos a algumas bactérias fixadoras de nitrogênio uma vez que 93 diferentes organismos foram encontrados. Ao contrário, os homólogos ao gene *draG* são ubíquos na natureza e a análise indicou 1.405 organismos, distribuídos nos três domínios de vida, incluindo os vírus. Finalmente, os resultados mostram que: os homólogos a enzima DraT estão restritos a poucos organismos fixadores de nitrogênio e que os homólogos a enzima DraG são comuns na natureza. Análise de vizinhança gênica sugere que proteínas homólogas a DraG, presentes em organismos não fixadores de nitrogênio, podem estar relacionadas com vias de degradação e/ou biossintese de nucleotídeos e/ou moléculas relacionadas.

Palavras-chave: Bioinformática, Fixação Biológica de Nitrogênio, ADP-ribolisalação, Dinitrogenase Redutase ADP-ribosil transferase, Dinitrogenase Redutase Glicohidrolase.

ABSTRACT

Nitrogenase is the enzyme responsible for the process of Biological Nitrogen Fixation (BNF), in which dinitrogen gas is reduced to ammonia by microorganisms. In some nitrogen-fixing bacteria such as Azospirillum brasilense, the enzyme activity is regulated at the post-translational level by ADP-ribosylation. The dinitrogenase reductase ADP-ribosyl transferase (DraT) enzyme catalyzes the ADP-ribosylation and then inactivates the nitrogenase. On the other hand, glicohidrolase dinitrogenase reductase (DraG) removes the modifier group and reactivates the nitrogenase. This work identified proteins similar to DraT and DraG enzymes, stored at the GenBank database, using bioinformatics software, especially BlastP and CD-HIT. The analysis revealed: the draT homologous genes are restricted to some nitrogen-fixing bacteria as 93 different organisms were found. In other hand, the draG homologous genes are ubiquitous in nature and the analysis indicated 1,405 organisms distributed into three different life domains, including viruses. Finally, the results show: DraT enzyme homologous are restricted to few nitrogen fixing organisms and, DraG enzyme homologous are common in nature. Gene neighborhood analysis suggests that DraG protein homologous, presented in non nitrogen-fixing organisms, may be related to degradation pathways and/or biosynthesis of nucleotides and/or related molecules.

Keywords: Bioinformatics, Biological Nitrogen Fixation, ADP-ribosylation, Dinitrogenase Reductase ADP-ribosiltransferase, Dinitrogenase Reductase ADP-ribosil-hydrolase.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

NCBI – National Center for Biotechnology Information

FeMo-co – Cofator Ferro Molibdênio

BLAST – Basic Local Alignment Search Tool

DraG – Dinitrogenase Redutase Glicohydrolase

DraT – Dinitrogenase Redutase ADP-ribosil Transferase

PDB – Protein Data Bank

NifH – Proteína Fe ou dinitrogenase redutase

NifDK – Proteína MoFe ou dinitrogenase

ARH – ADP-ribosil hidrolases

XML – Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

Desde a década de 80, a Bioinformática começou a ser objeto de estudo e após a finalização do projeto genoma humano, em 2003, tornou-se uma área de destaque para pesquisadores (HSU, 2006; p. vi). Devido ao desenvolvimento das tecnologias de seqüenciamento, a quantidade de dados biológicos gerados alcançou altos níveis o que impossibilitou as análises manuais, permitindo o surgimento e o crescimento da área da ciência conhecida como Bioinformática (PROSDOCIMI, 2007).

Define-se que a Bioinformática é uma abordagem computacional para a gestão e análise de informação biomédica, utilizada cada vez mais como um componente auxiliador para a investigação tanto em ambientes acadêmicos quanto industriais (NCBI, 2002).

A Bioinformática ou biologia computacional é uma ciência interdisciplinar de interpretação de dados biológicos (RAZA, 2011). É constituída por variadas ciências e possui sua essência em um modelo interdisciplinar (FIGURA 1), (BAYAT, 2002). Pode ser dividida em duas vertentes: a bioinformática tradicional e a bioinformática estrutural. A primeira está relacionada as seqüências de aminoácidos e nucleotídeos, e a segunda aborda questões relacionadas a estrutura tridimensional e modelagem molecular (VERLI, 2014).

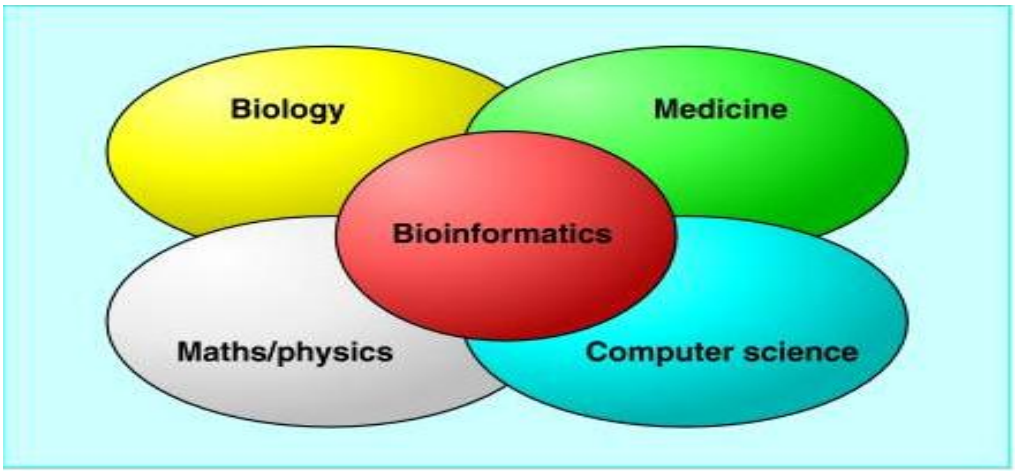

FIGURA 1 - INTERAÇÃO DE DISCIPLINAS QUE ENGLOBAM A BIOINFORMÁTICA
FONTE: BAYAT (2002)

Os autores Loman e Watson (2013), afirmam que os computadores são componentes essenciais da pesquisa biológica moderna e que os cientistas da área estão sendo convidados a adotar novas habilidades computacionais. Visto que, geralmente os dados biológicos são depositados em bancos de dados, os quais estão disponíveis na World Wide Web, além das ferramentas para análises de Bioinformática (HSU, 2006; p. vi).

E o rápido desenvolvimento da Bioinformática exige de seu praticante uma constante atenção à novas abordagens, métodos e tendências. As predições a partir de sequências são um exemplo destas novas abordagens, assim como a análise de genes vizinhos (VERLI, 2014).

1.2 JUSTIFICATIVA

O Nucleo de Fixação Biológica de Nitrogênio da UFPR possui como um de seus objetos de estudo a regulação e controle pós-traducional da enzima nitrogenase por ADP-ribosilação assim como também as enzimas envolvidas no processo, como as enzimas DraT e DraG. Este trabalho tem como objetivo realizar uma análise de bioinformática dos genes *draT* e *draG* a fim de revelar a distribuição de homólogos e uma análise de vizinhança destes genes em diferentes organismos.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Análise bioinformática dos genes *draT* e *draG*.

1.3.2 Objetivos Específicos

- Identificar proteínas similares às enzimas DraT e DraG de *Azospirillum brasilense* depositadas no banco de dados GenBank.

- Determinar quais organismos codificam para proteínas similares às enzimas DraT e DraG de *Azospirillum brasilense*.

- Verificar se os organismos que codificam a enzima DraT também codificam a enzima DraG e os genes do cluster *nif* mínimos para a FBN.

- Analisar a vizinhança dos genes *draT e draG* nos diferentes organismos.

2 REVISÃO DE LITERATURA

2.1 FIXAÇÃO BIOLÓGICA DE NITROGÊNIO

O gás dinitrogênio (N_2) é o gás em maior quantidade na atmosfera terrestre, constituindo cerca de 80% da mesma, seres eucariotos não possuem a capacidade de utilizar o nitrogênio atmosférico de forma direta (POSTGATE, 1982). O nitrogênio atmosférico pode ser fixado em formas utilizáveis pelos eucariotos principalmente através da produção de fertilizantes nitrogenados ou através de microorganismos que convertem o nitrogênio atmosférico (N_2) a amônio (NH_3) (PEDROSA, 1987).

Em consequência de que os fertilizantes nitrogenados são caros e o seu uso em excesso torna-se um agente poluente ao meio ambiente a Fixação Biológica de Nitrogênio (FBN) revelou-se uma alternativa para a diminuição de gastos na produção agrícola, pois bactérias diazotróficas se associam as plantas e converter o nitrogênio atmosférico a amônio aumentando a fertilidade do solo (PEDROSA, 1987).

Para que o N_2 fique em uma forma metabolicamente utilizável ele precisa ser transformado em NH_3 através do complexo enzimático denominado nitrogenase (POSTGATE, 1982). Esta fixação biológica é desempenhada somente por bactérias denominadas de diazotróficas (BURRIS, 1991) distribuídas entre os domínios Archaea e Bacteria (DIXON e KAHN, 2004).

A FBN possui a importante função de manutenção do ciclo do nitrogênio na biosfera (POSTGATE, 1982) e para os eucariotos a relevância deste processo encontra-se na dependência para síntese de biomoléculas (ácidos nucléicos e proteínas) que utiliza o amônio resultante da FBN (BURRIS, 1991).

2.2 NITROGENASE

A nitrogenase é um complexo enzimático que é responsável pela reação de catálise no processo de Fixação Biológica de Nitrogênio, no qual o gás dinitrogênio é convertido em íons amônio. Esra enzima é composta por duas sub-unidades: o

dímero menor (γ_2) denominado de dinitrogenase redutase, proteína Fe ou NifH e a dinitrogenase, proteína MoFe ou NifDK ($\alpha_2\beta_2$) portadora do sítio ativo da enzima (FIGURA 2) (RAYMOND et al., 2004; DIXON e KAHN, 2004).

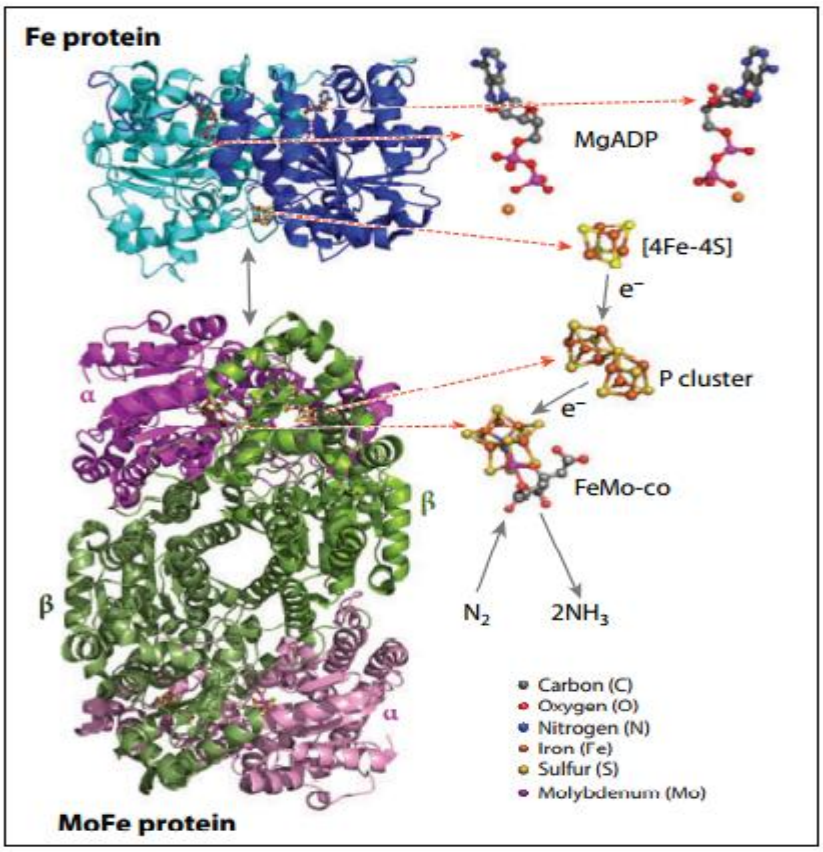

FIGURA 2 – MODELO ESTRUTURAL DO COMPLEXO NITROGENASE DE *Azotobacter vinelandii*
FONTE: SEEFELDT *et al*. (2009)

Em uma reação com alto gasto energético, a proteína Fe doa elétrons de forma dependente de ATP, para a proteína MoFe e para cada elétron transportado são consumidas 2 moléculas de ATP (SEEFELDT et al., 2009). Portanto, para a redução de cada N_2 para $2NH_3$, são consumidas 16 moléculas de ATP, como mostra a reação abaixo (EADY, 1986; POSTGATE, 1982):

$$N_2 + 10H^+ + 8e^- + 16Mg.ATP \xrightarrow{\text{Nitrogenase}} 2NH_4^+ + H_2 + 16Mg.ADP + 16Pi$$

2.3 ORGANISMOS DIAZOTRÓFICOS

Segundo POSTAGE (1982), algumas espécies de organismos procariontes, denominados de diazotrofos, que podem ser encontrados nos domínios Archaea e Bactéria, são capazes de fixar nitrogênio.

Estes organismos podem viver em vida livre ou em associação com outros organismos. Neste último caso, enquanto a bactéria fornece nitrogênio para a planta, esta fornece em troca uma fonte de carbono. Entretanto, os organismos diazotróficos de vida livre, fixam nitrogênio para seu próprio uso sem qualquer associação a demais organismos de (PELCZAR *et. al*, 1996).

A reação de incorporação do nitrogênio pelas bactérias diazotróficas é catalisada pelo complexo enzimático da enzima Nitrogenase e consiste na conversão do nitrogênio gasoso (N_2) na sua forma mais reduzida, como íons de amônio $(NH_4)^+$, que serão utilizados pelos seres vivos para a biossíntese de seus compostos nitrogenados (POSTGATE, 1982).

2.4 CLUSTER *nif*

Diversos genes estão envolvidos no processo de FBN e estes estão organizados em grupos ou “clusters”. O cluster *nif* (“nitrogen fixation”) possui genes requeridos para a estrutura, biossíntese e regulação da transcrição da nitrogenase (DIXON e KAHN, 2004).

Os genes *nif* estão dispostos em várias unidades transcricionais de regulação. Além dos genes estruturais da Nitrogenase (*nifHDK*), essas unidades também possuem os genes que controlam as proteínas de transporte de elétrons pois codificam para produtos envolvidos na biossíntese do co-fator Fe-Mo da nitrogenase, e certos genes reguladores da transcrição (DIXON e KAHN, 2004).

Desde sua primeira descrição em *Klebsiella pneumoniae* (CANNON *et al*.,1974) o cluster *nif* vem sendo objeto de estudo, e Dixon e colaboradores, identificaram (no mesmo organismo) 21 genes *nif* em uma região de 20kb do genoma (DIXON *et. al*., 1986).

Porém em estudos recentes Dos Santos *et. al.* (2012), afirma que a maioria dos

organismos fixadores possuem um mínimo de seis genes conservados, sendo eles: *nifH, nifD, nifK, nifE, nifN* e *nifB* (FIGURA 3).

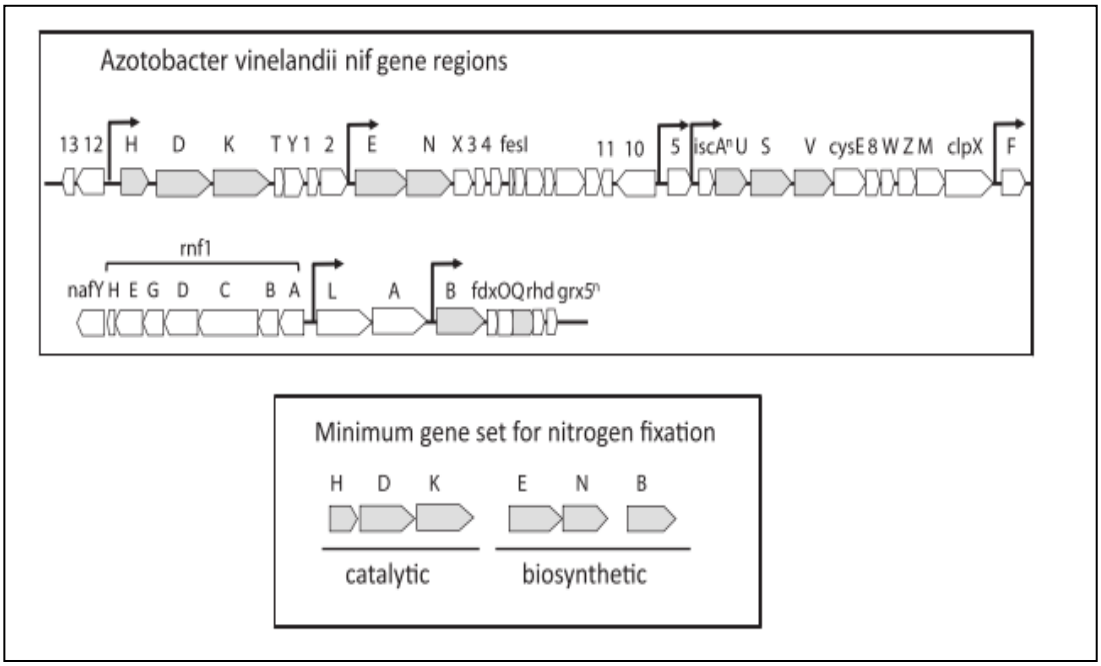

FIGURA 3 – GENES ENVOLVIDOS NA FIXAÇÃO DE NITROGÊNIO EM *A. vinelandii*
FONTE: DOS SANTOS *et al*. (2012)

2.5 REGULAÇÃO PÓS-TRADUCIONAL DA NITROGENASE

Em *Rhodospirillum rubrum* (LOWERY et al., 1986), *Azospirillum brasilense* (HARTMANN et al., 1986), *Rhodobacter capsulatus* (JOUANNEAU et al., 1983) e em alguns outros organismos diazotrofos a atividade da enzima nitrogenase é regulada de forma pós-traducional. Esta regulação ocorre por ADP-ribosilação de um resíduo de arginina em uma das subunidades da proteína Fe, catalisada pela enzima DraT. Este processo resulta na inativação da proteína Fe. Em contra partida, a ADP-ribosilação da proteína Fe pode ser revertida pela DraG ativando, desta maneira, a nitrogenase (ZHANG et al., 1997). Este fenômeno é conhecido como “swich-off/swich-on” desligamento e religamento da nitrogenase (ZUMFT e CASTILLO, 1978).

As enzimas DraT e DraG foram caracterizadas primeiramente em *Rhodospirillum rubrum* e são codificadas pelo operon *draTG*, o qual se localiza a montante dos genes estruturais da nitrogenase o *nifHDK* (LIANG *et. al.,* 1991).

Zhang e colaboradores (1997), observaram que as atividades de DraT e DraG são reguladas de maneira inversa *in vivo*. Como resposta a um estímulo negativo (presença de íons amônio ou diminuição da energia celular), DraT torna-se ativa e nestas condições, ocorre a inativação da nitrogenase por ADP-ribosilação da proteína Fe. Porém, quando este estímulo negativo é excluído, DraT é inativada e DraG ativada promovendo desta forma a remoção do grupo ADP-ribosil e por consequência a restauração da a atividade da nitrogenase (HUERGO et.al., 2012; ZHANG et.al., 1997) (FIGURA 4).

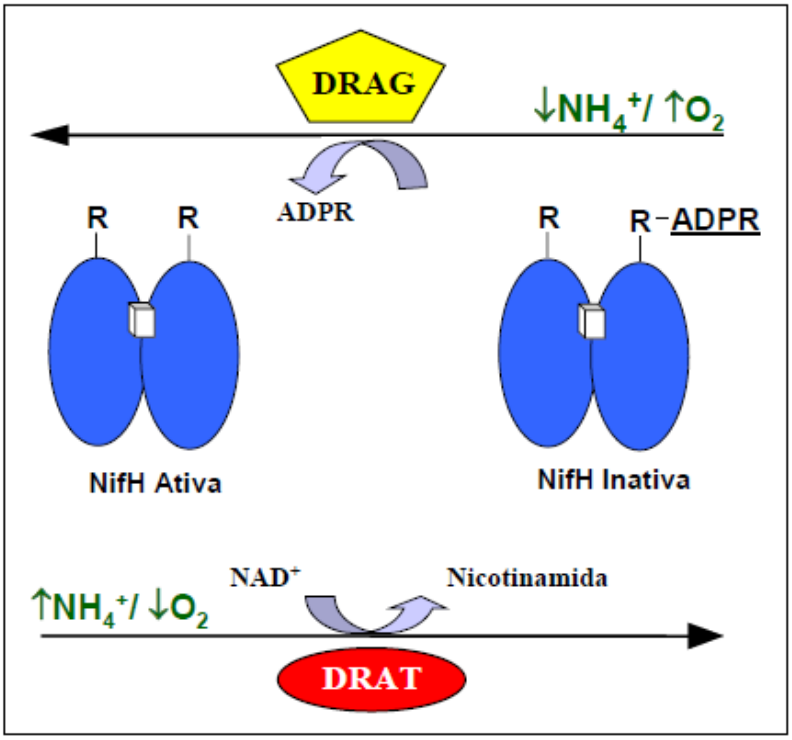

FIGURA 4 – ESQUEMA DA INATIVAÇÃO REVERSÍVEL DE NIFH POR ADP-RIBOSILAÇÃO
FONTE: HUERGO (2006)
Nota: Os dímeros da proteína NifH estão representados em azul, a letra R indica o resíduo de arginina 101, R-ADPR indica o resíduo Arg101 modificado por ADP-ribose. Em resposta a presença de íons amônio ou oxigênio, DraT é ativada para catalisar a ADP-ribosilação da proteína Fe, inativando a nitrogenase; quando o amônio é consumido ou os níveis de energia se restabelecem, DraG remove o grupo ADP-ribosil, reativando a nitrogenase.

2.6 DraT

DraT é um monômero de 30 kDa, que catalisa a transferência de ADP-ribosil de NAD^+ para a cadeia lateral de R101 de uma das subunidades proteína Fe (LOWERY et al., 1986). Sua principal função conhecida é a capacidade de modificar

a proteína Fe, pois é o seu único substrato conhecido. (LOWERY e LUDDEN, 1989). Segundo Nordlund e Högbom (2013), nenhum modelo estrutural *in silico* pode ser obtido, pois a DraT é instável *in vitro*, o que dificulta sua cristalização.

A atividade bioquímica de DraT é semelhante a um certo número de ADP-ribosil-transferases de toxinas bacterianas para as quais a estrutura é conhecida, tal como a toxina iota (TSUGE *et. al*., 2003), a toxina da difteria (BEEL e EISENBERG, 1996) a toxina Certhrax (VISSCHEDYK *et. al*., 2012) e como parte da toxina cólera (HOTTIGER *et. al*., 2010).

2.7 DraG

Com estrutura cristalográfica determinada, a DraG é um monômero de 32 kDa, primeiramente do *R. rubrum* (SAARI *et. al.,* 1984, 1986). Esta enzima quebra da ligação glicosídica entre o grupo ADP-ribosil que está ligado ao resíduo de arginina (LJUNGSTROM *et. al.,* 1989; SAARI *et. al.*, 1984) (FIGURA 5) e ao contrário da DraT, DraG catalisa a remoção da porção ADP-ribosil (POPE et al. , 1986).

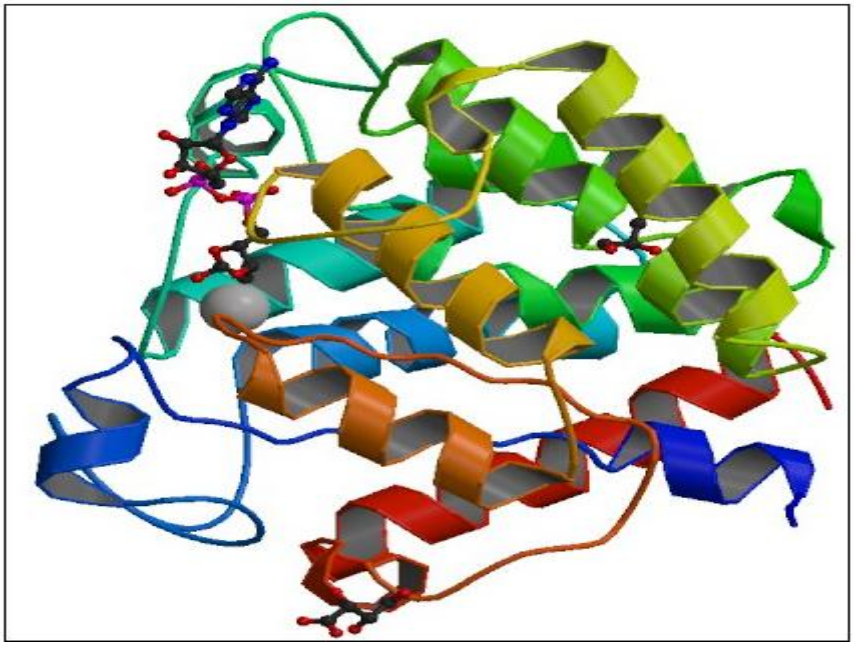

FIGURA 5 – ESTRUTURA DRAG DE *Rhodospirillum rubrum*

FONTE: PDB (2013)

DraG tornou-se um modelo de ADP-ribosilhidrolases (ARH) específicas de arginina, cuja função biológica está bem estabelecida e tem um modelo de mecanismo catalítico proposto (BERTHOLD et al., 2009).

A enzima DraG é homologa à outras proteínas disponíveis no PDB, e na maioria dos casos de homologia observa-se que a estrutura geral é muito semelhante, pois a proteína é constituída por 15-19 alfa-hélices e o centro ativo é localizado dentro de uma fenda hidrofílica rodeada por quatro regiões de voltas internas (LI *et. al*., 2009). A proteína ARH3 humana possui uma estrutura tridimensional muito similar a enzima DraG, apesar da baixa similaridade na seqüência de aminoácidos (FIGURA 6).

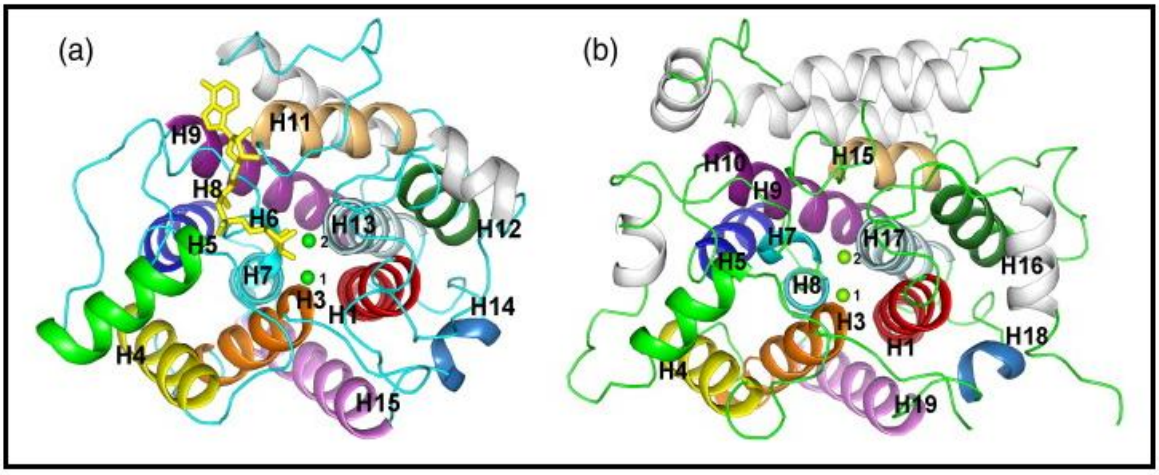

FIGURA 6 – ESTRUTURA GERAL DE DraG DE *A. brasilense* E COMPARAÇÃO COM A ARH3 HUMANA. (a) Proteína DraG identificando as α-hélices H1-H13. (b) Proteína ARH3 Humana e suas 13 α-hélices conservadas. Os dois íons magnésio são mostrados como esferas verdes (marcadas 1 e 2)

FONTE: Adaptado de Li *et. al*. (2009)

3 MATERIAIS E MÉTODOS

3.1 MATERIAIS

Os materiais descritos nas seções a seguir correspondem às ferramentas utilizadas e que auxiliaram o desenvolvimento desta pesquisa.

3.1.1 Banco de Dados

Os genomas completos e inúmeras outras informações utilizadas neste trabalho foram retiradas do *National Center for Biotechnology Information* (NCBI), pois segundo Benson (2008), o NCBI é o maior banco de dados da atualidade, em termos de informações sobre sequências. Ele engloba aproximadamente 35 bancos de dados integrando informações da maioria das bases de seqüências de DNA e proteínas juntamente com as de taxonomia, genomas, mapeamento, estruturas de proteínas e informações de domínios estruturais além de referencias bibliográficas biomédicas (PubMed).

As seqüências de aminoácidos das oito proteínas pesquisadas foram retiradas do Uniprot, o qual é o banco de dados que armazena informações funcionais sobre as proteínas possuindo anotações precisas, consistentes e ricas. É constituído por duas seções: uma com registros anotados manualmente e análise computacional (UniProtKB / Swiss-Prot) e a outra sessão com registros computacionais esperando por anotação manual (UniProtKB / TrEMBL) (UniProt, 2015).

O STRING é um banco de dados dedicado a interações proteína-proteína, incluindo informações de interações físicas e funcionais (VON MERING *et.al*., 2003). Este software foi utilizado na versão 9.1.

3.1.2 Linguagem de Programação

A implementação computacional realizada neste trabalho e denominada de

Blaster foi desenvolvida na linguagem de programação JAVA. Esta linguagem foi escolhida devido a existência de uma biblioteca nomeada BioJava.

Com a característica de código aberto, a biblioteca BioJava é dedicada a disponibilizar ferramentas para o processamento de dados biológicos sendo útil para automatizar tarefas relacionadas à Bioinformática (PRLIC *et. al*., 2012).

3.1.3 Softwares

O *Basic Local Alignment Search Tool* (BLAST) é um software para localização de regiões de similaridade local entre sequências. Este programa compara as sequências de nucleotídeos ou de aminoácidos e calcula a significância estatística dos resultados. O BLAST pode ser utilizado para inferir relações funcionais e evolutivas entre sequências (ALTSCHUL *et.al.*,1990). Contudo, o BLASTP foi projetado para encontrar regiões locais de similaridade em base dados de proteínas (NCBI, 2014). Nesta pesquisa o BLASTP (*release 2.2.28*), foi utilizado com o objetivo de verificar a existência de proteínas homólogas em outros organismos.

Para realizar a análise de vizinhança foi utilizado o software CD-HIT, este é um programa que realiza agrupamentos a partir dos dados de entrada (sequências de proteínas ou de nucleotídeos). Esta formação de grupos ocorre de forma rápida e independente da quantidade de dados. Basicamente, as sequências são classificadas em ordem decrescente de comprimento. A sequência considerada como maior se torna a representante de um grupo. Em seguida, cada sequência restante é comparada com os representantes dos agrupamentos existentes. Se a semelhança com qualquer representante está acima de um determinado limite, este será agrupado, caso contrário um novo grupo será definido com essa sequência considerada como representante (LI et. al., 2012).

Os parâmetros para a execução dos algoritmos das ferramentas BLASTP e CD-HIT estão apresentados na (TABELA 1).

TABELA 1 - PARÂMETROS UTILIZADOS PARA A EXECUÇÃO DO BLAST E CD-HIT

Parâmetros BLAST	
Database	Non-redundant protein sequences (nr)
Program Selection Algorithm	Blastp (protein-protein BLAST)
Max target sequences	20000
Matrix	BLOSUM62
Parâmetros CD-HIT	
Sequence identity cut-off	0.3
G: use global sequence identity	YES
g: sequence is clustered to the best cluster that meet the thershold	YES
b: bandwidth of alignment	20

O Microsoft Office Excel (versão 2007) foi utilizado como ferramenta de armazenamento, organização e manipulação dos dados gerados, pois é um software constituído de planilhas eletrônicas de fácil usabilidade (MICROSOFT, 2014).

3.2 MÉTODOS

A metodologia deste trabalho, para melhor entendimento, está divida em quatro fases, sendo elas: obtenção dos dados, filtragem, agrupamento e organização dos resultados e análise de vizinhança, as quais serão descritas a seguir (FIGURA 7).

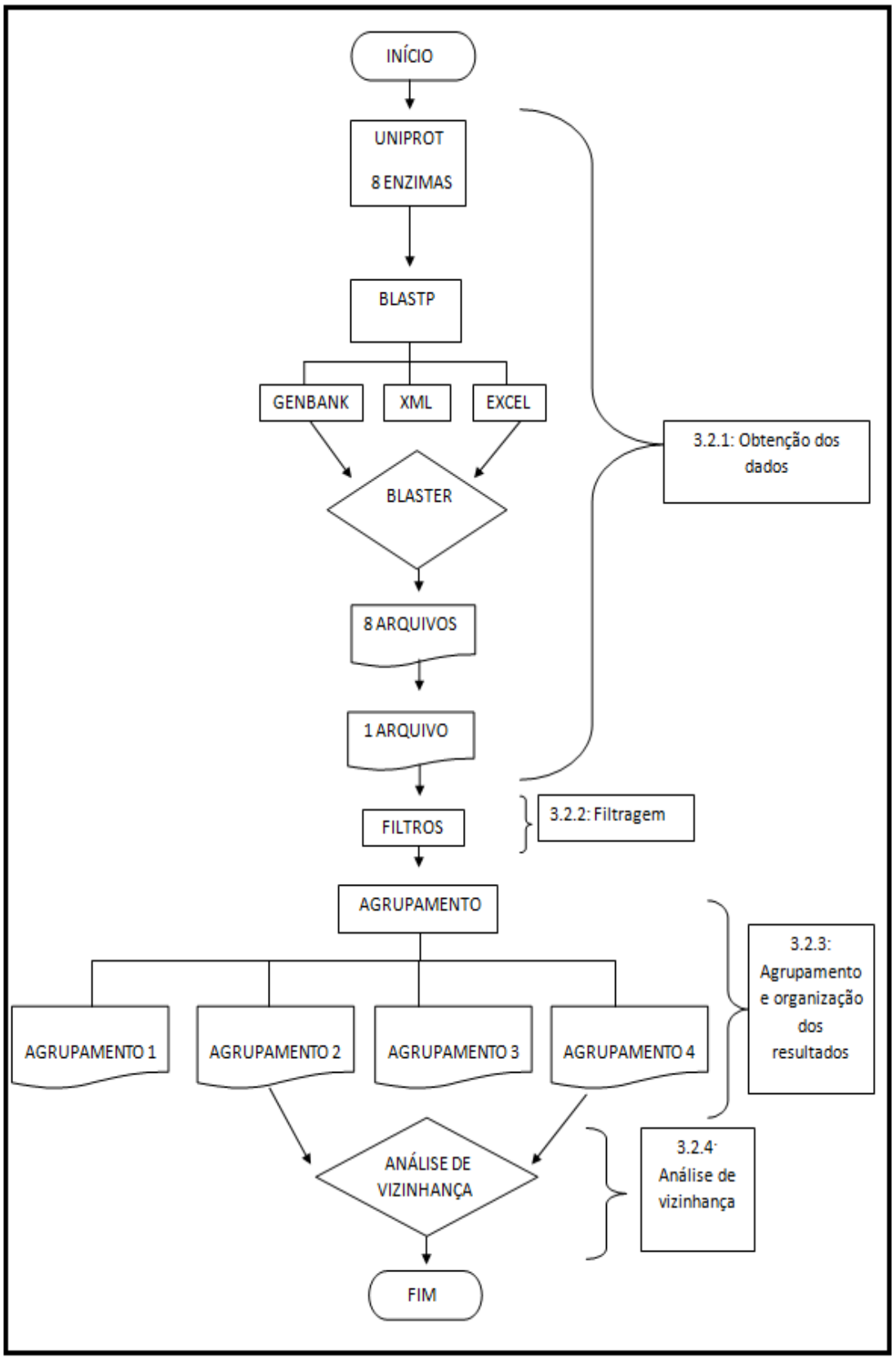

FIGURA 7 – FLUXOGRAMA REFERENTE À METODOLOGIA

FONTE: A autora (2014)

3.2.1 OBTENÇÃO DOS DADOS

Na etapa inicial de obtenção dos dados realizou-se a localização e extração das seqüências de aminoácidos de oito proteínas depositadas no banco de dados UniProt do organismo *A. brasilense* sp.245 (APÊNDICE 1), sendo elas: DraT (AZOBR_70132), DraG (AZOBR_70133), NifN (AZOBR_70121), NifB (AZOBR_70147), NifE (AZOBR_70122), NifK (AZOBR_70126), NifH (AZOBR_70128) e NifD (AZOBR_70127). O organismo *A. brasilense sp.245* é uma bactéria diazotrófica e é objeto de estudo do Grupo de Fixação de Nitrogênio da Universidade Federal do Paraná, assim como a *A. brasilense* FP2 e Sp7.

Após a obtenção das seqüências, foram realizados oito alinhamentos locais (em 08/13 - *release 2.2.28*), utilizando o Blastp com a finalidade de encontrar proteínas homólogas com os dados de entrada *(query).*

Os oito resultados das pesquisas foram armazenados cada um deles em três formatos diferentes: o Hit table (formato em Excel), o alinhamento (formato em XML) e o GENBANK (formato em GBK), portanto neste momento a pesquisa total possui 24 arquivos.

Informações importantes como: gi, porcentagem de identidade e de positivos, gaps, E-value; encontram-se somente no formato Hit table. No formato GENBANK são apresentados outros dados importantes, como: *locus*, *acession*, *version*, *definition*, organismo, dentre outros. O formato em XML contém o alinhamento da busca realizada. Portanto, os formatos com dados de maior interesse para a pesquisa encontram-se nos formatos Hit table e no GENBANK e a pesquisa totaliza 16 arquivos. Mesmo com a redução do número de arquivos, o manuseio destes tornou-se uma dificuldade.

Diante disto foi necessário o desenvolvimento de uma aplicação computacional com o objetivo de unir os dois tipos de formatos em apenas um arquivo para cada busca realizada. Desta forma elaborou-se uma aplicação denominada de Blaster, a qual teve seu código fonte desenvolvida na linguagem de programação JAVA (APÊNDICE 4). O Blaster possui a função de unir dois arquivos de formatos diferentes (Hit table e GENBANK), mas estes deverão ser oriundos de uma mesma pesquisa. Na figura 8, pode-se visualizar a tela única do software que possui dois campos de entrada e um para a saída dos dados.

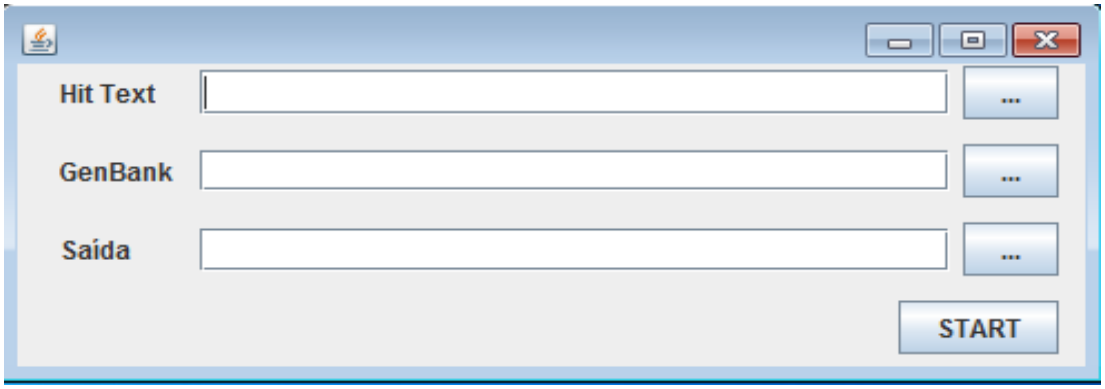

FIGURA 8 – CAPTURA DE TELA DO SOFTWARE BLASTER.

FONTE: A autora (2014).

Com o auxílio da ferramenta Blaster, obteve-se oito arquivos um para cada proteína.

Em virtude da melhor manipulação dos dados, os 8 arquivos foram concatenados transformando-se em um único arquivo, porém com um universo com mais de 30.000 linhas que caracterizam os achados da pesquisa.

A próxima etapa da metodologia se baseia em reduzir a quantidade de dados através do uso de filtros que serão explicados a seguir.

3.2.2 FILTRAGEM

Nesta etapa denominada de filtragem, os dados da pesquisa foram selecionados a partir de filtros específicos, sendo eles: organismos não cultivados, porcentagem de identidade e valor de E-value.

3.2.2.1 Organismos não cultivados

Neste primeiro filtro os organismos não cultivados são excluídos, pois não possuem relevância para o estudo. Em virtude deste filtro, a quantidade de achados passou de mais de 30.000 para 19.884 achados (hits).

3.2.2.2 Porcentagem de Identidade

O valor de porcentagem de identidade de um alinhamento é expressivo, pois é o número que indica a quantidade de nucleotídeos ou aminoácidos pareados (FASSLER e COOPER, 2011).

No caso desta pesquisa, foram considerados apenas os valores de identidade (percentual) iguais e/ou superiores a 20 para todas as enzimas, pois segundo EMBL e LION (1999), percentuais de identidades menores que 20 - 35% podem gerar um falso positivo, tornando uma tarefa difícil afirmar que tais sequências são relacionadas.

3.2.2.3 E-value

O valor E-value é um indicador do grau de significância estatística do resultado de uma pesquisa, pois mesmo considerando um alto *score* no alinhamento, nem sempre é possível inferir homologia. Em conseqüência da aleatoriedade do processo de alinhamento, o E-value fornece credibilidade ao achado, revelando a probabilidade de que um hit ocorreu ou não ao acaso. Evidentemente, quanto menor o valor mais significativo será o resultado encontrado (FASSLER e COOPER, 2011).

Os valores de E-value foram utilizados como filtros (TABELA 2), para as enzimas. Para DraT, NifH, NifD, NifE, NifK,NifN, NifB foram considerados os valores de E-value igual e/ou menor de $1x10^{-20}$. Para a enzima DraG foi considerado um valor de E-value igual e/ou menor de $1x10^{-3}$, pois é o valor de E-value entre a proteína DraG de *A. brasilense* e a proteína ARH3 humana, tratadas nesta pesquisa. Apesar do alto valor de E-value, sabe-se que DraG e ARH3 tem alta homologia/similaridade estrutural (LI *et. al.*, 2009).

TABELA 2 - VALOR DE CORTE DE E-VALUE PARA CADA PROTEÍNA

Proteína	*E-value*
DraG	$<=1x10^{-3}$
DraT	
NifH	
NifD	

NifE	<=1x10^{-20}
NifK	
NifN	
NifB	

Após a filtragem dos dados, obteve-se um universo com 10.961 achados (*hits*). Pode-se sugerir que a proteína NifH é altamente conservada (nesta pesquisa), pois se manteve com o mesmo número de *hits* após o processo de filtragem (TABELA 3).

TABELA 3 - NÚMERO DE HITS ENCONTRADOS X NÚMERO DE HITS FILTRADOS

Enzima	*Hits encontrados no Blastp* **excluindo os organismos não cultivaveis*	*Hits com os filtros*
DraT	115	106
DraG	2980	2031
NifD	2840	1698
NifE	3376	1546
NifH	2598	2598
NifK	2662	1174
NifN	1730	1128
NifB	3583	679
Total	19884	10961

3.2.3 AGRUPAMENTO E ORGANIZAÇÃO DOS RESULTADOS

Tratamos a terceira fase da metodologia com quatro agrupamentos distintos em tabelas distintas no software EXCEL.

O 1º Agrupamento tem como guia a enzima DraT, ou seja, todos os organismos que possuem a enzima DraT e sem duplicidade de organismos, constam nesta tabela, independente de se ter ou não as outras enzimas (DraG e o cluster *nif* mínimo). Estes dados foram armazenados na tabela denominada de Agrupamento 1.

Para o 2º Agrupamento foi utilizado o mesmo método do 1º agrupamento, porém permaneceram apenas os organismos com genoma completo. Estes dados foram armazenados na tabela denominada de Agrupamento 2.

O 3º Agrupamento contém todos os organismos que possuem a enzima DraG, com genomas completos e não completos. Estes dados foram armazenados na tabela denominada de Agrupamento 3.

Para o 4º Agrupamento foi utilizado o mesmo método do 3º agrupamento, porém permaneceram apenas os organismos com genoma completo e com a presença da proteína DraG (sem as outras proteínas da pesquisa), ou seja, organismos não fixadores de nitrogênio. Estes dados foram armazenados na tabela denominada de Agrupamento 4.

3.2.4 ANÁLISE DE VIZINHANÇA

A investigação dos genes vizinhos pode sugerir funções conhecidas anteriormente pois, segundo LESK (2008), em muitos genomas procarióticos existem regiões denominadas de *operons*, que são genes adjacentes que são transcritos em uma única molécula de RNA mensageiro, sob o mesmo controle transcricional. Em bactérias, os genes contidos em um mesmo *operon* podem estar funcionalmente conectados. Portanto, se as funções dos produtos de genes são conhecidas estas informações podem sugerir funções e produtos de genes desconhecidos (Dandekar *et. al*., 1998).

Os conjuntos de dados escolhidos para realizar esta análise de vizinhança foram os Agrupamentos 2 e 4, referentes as proteínas DraT e DraG, respectivamente, pois estes agrupamentos são compostos por organismos com genomas completos.

Foram considerados como genes vizinhos os genes contidos em uma janela de 5 genes anotados a montante (*upstream*) e 5 genes a jusante (*downstream*) da proteína de interesse, no caso DraT ou DraG.

Na tabela 4, é possível observar que para ambas as proteínas, a ocorrência do gene a ser estudado pode apresentar-se (em alguns casos) mais de uma cópia no mesmo organismo, ou seja, um dado organismo poderá apresentar mais de uma cópia da proteína em seu genoma, possivelmente indicando duplicação da proteína. Para compor o conjunto de dados foram retirados os vizinhos que continham apenas o número de Gene ID, ou seja, não continham dados de proteínas, pois quando anotados não se tratavam de uma região codificadora. Portanto, a análise de vizinhança para o gene *draT* contou com 569 genes e para o gene *draG* com 6.652 genes.

TABELA 4 - PANORAMA DA ESTRATÉGIA DE ANÁLISE DE VIZINHANÇA

Gene	*Ocorrência da proteína X número de organismos*	*Número de Gene ID (retirados)*	*Número de Genes Vizinhos*
draT	58/51	11	569
draG	682/383	168	6652

Com a finalidade de compreender melhor a vizinhança dos genes *draT* e *draG*, utilizou-se a metodologia de agrupar os genes vizinhos em três diferentes arquivos distintos, com diferentes janelas de vizinhanças, sendo elas: de -1 a +1, de -3 a +3 e de -5 a +5, denominadas de grupos de estratégia.

Para realizar esta análise, utilizou-se o software CD-HIT com a finalidade agrupar proteínas semelhantes. Para cada proteína foram realizadas 3 pesquisas devido os grupos de estratégia estabelecidos, totalizando 6 pesquisas (TABELA 5).

TABELA 5 - QUANTIDADE DE GENES POR GRUPOS DE ESTRATÉGIA APLICADA PARA A ANÁLISE DE VIZINHANÇA

Gene	-1 a +1	-3 a +3	-5 a +5
draT	115	338	569
draG	1337	4000	6652

Após a geração dos dados foi necessário utilizar critérios de exclusão, a saber: a) os grupos com apenas uma seqüência e b) os grupos com mais de uma seqüência, porém com organismos da mesma espécie.

Na tabela 6, é possível observar a quantidade de grupos formados, grupos com mais de uma seqüência, grupos com somente uma seqüência e grupos com mais de uma seqüência e com diferentes espécies. Obviamente, o grupo de estratégia de -5 a +5 se sobrepõe aos demais, devido à maior quantidade de genes contidos neste agrupamento. A última coluna contém a quantidade de grupos válidos para este trabalho.

TABELA 6 - PANORAMA DOS GRUPOS GERADOS PELO SOFTWARE CD-HIT

Grupos de estratégia	*Quantidade de grupos gerados*	*Quantidade de grupos com mais de uma seqüência gênica*	*Quantidade de grupos com uma seqüência gênica*	*Quantidade de grupos com mais de uma seqüência gênica e com organismos de espécies diferentes*
		draT		
DE -5 A +5	271	63	208	12
DE -3 A +3	148	33	115	8
DE -1 A +1	43	10	33	4
		draG		
DE -5 A +5	3636	930	2706	115
DE -3 A +3	2179	247	1932	146
DE -1 A +1	733	164	570	40

4 RESULTADOS E DISCUSSÃO

Neste trabalho o primeiro resultado obtido foi o desenvolvimento da ferramenta Blaster. Com esta aplicação foi possível manipular e unificar informações necessárias de maneira rápida e eficaz. E após a identificação de genes homólogos, realização da clusterização destes genes e análise de vizinhança para os genes *draT* e *draG*, obtiveram-se os seguintes resultados.

A análise realizada referente ao Agrupamento 1, a qual contou com a sequência da enzima DraT em todos os achados, revelou que esta proteína encontra-se restrita ao domínio Bacteria. Foi identificada em 93 organismos diferentes (organismos com genomas completos e não completos) a DraT, sendo predominante no filo *Proteobacteria* (classes *Alpha*, *Beta*, *Delta* e *Gamma*). Entretanto, também está distribuída nos filos *Chrysiogenetes*, *Deferribacteres* e *Verrucomicobria* (classes *Opitutae* e *Verrucomicrobiae*) (FIGURA 9).

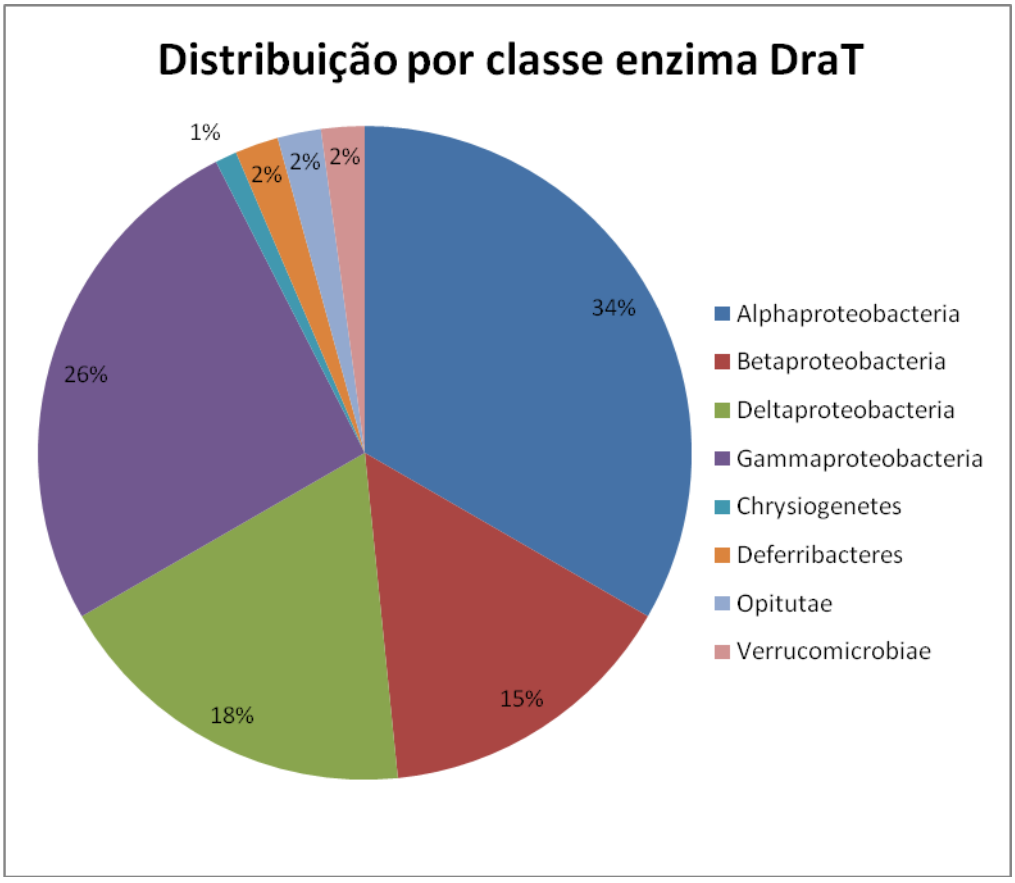

FIGURA 9 – DISTRIBUIÇÃO TAXONÔMICA POR CLASSE DA ENZIMA DraT
FONTE: A autora (2014)

A segunda análise realizada (Agrupamento 2) revelou que a DraT é restrita à bactérias diazotróficas, pois os genomas completos tratados nesta pesquisa que contém o gene *draT* também possuem os genes *nif (nifH*, *nifE*, *nifD*, *nifN*, *nifB* e *nifK*) e o gene *draG* distribuídos em 51 organismos diferentes descritos na tabela 7 e no Apêndice 2.

TABELA 7 - DISTRIBUIÇÃO TAXONÔMICA DOS ORGANISMOS QUE POSSUEM OS 8 GENES

Filo	**Classe**	**Organismo**
Proteobacteria	**Alphaproteobacteria**	*Azospirillum brasilense Sp245***
		*Azospirillum lipoferum 4B***
		Azospirillum sp. B510
		Magnetococcus marinus MC-1
		Magnetospirillum gryphiswaldense MSR-1
		*Magnetospirillum magneticum AMB-1***
		*Rhodobacter capsulatus SB 1003***
		*Rhodobacter sphaeroides ATCC 17025***
		*Rhodopseudomonas palustris BisA53***
		*Rhodopseudomonas palustris BisB18***
		*Rhodopseudomonas palustris BisB5***
		*Rhodopseudomonas palustris CGA009***
		*Rhodopseudomonas palustris DX-1***
		*Rhodopseudomonas palustris HaA2***
		*Rhodopseudomonas palustris TIE-1***
		*Rhodospirillum rubrum ATCC 11170***
		Rhodospirillum photometricum DSM 122
		*Rhodospirillum rubrum F11***

Proteobacteria	**Betaproteobacteria**	*Azoarcus sp. BH72*
		*Azoarcus sp. KH32C***
		Candidatus Accumulibacter phosphatis clade IIA str. UW-1
		Dechloromonas aromatica RCB
		*Dechlorosoma suillum PS***
		*Rubrivivax gelatinosus IL144***
		Sideroxydans lithotrophicus ES-1
	Deltaproteobacteria	*Anaeromyxobacter sp. Fw109-5*
		Anaeromyxobacter sp. K
		Desulfobacterium autotrophicum HRM2
		Geobacter bemidjiensis Bem
		Geobacter daltonii FRC-32
		Geobacter lovleyi SZ
		*Geobacter metallireducens GS-15***
		Geobacter sp. M18
		Geobacter sp. M21
		*Geobacter sulfurreducens PCA***
		*Geobacter uraniireducens Rf4***
		Pelobacter carbinolicus DSM 2380
		*Pelobacter propionicus DSM 2379***
		*Acidithiobacillus ferrivorans SS3***
		*Acidithiobacillus ferrooxidans ATCC 53993***
		Allochromatium vinosum DSM 180
		Methylococcus capsulatus str. Bath

	Gammaproteobacteria	*Methylomonas methanica MC09***
		*Teredinibacter turnerae T7901***
		*Thiocystis violascens DSM 198***
		*Thioflavicoccus mobilis 8321***
		Tolumonas auensis DSM 9187
Chrysiogenetes	**Chrysiogenales**	*Desulfurispirillum indicum S5*
Deferribacteres	**Deferribacterales**	*Calditerrivibrio nitroreducens DSM 19672*
		Denitrovibrio acetiphilus DSM 12809
Verrucomicrobia	**Opitutae**	*Coraliomargarita akajimensis DSM 45221*

Legenda: ** possíveis novos organismos fixadores de nitrogênio.

Dos 51 organismos contendo a proteína DraT, 27 organismos (identificados com **) não foram mencionados como organismos diazotrofos por Dos Santos (2012) (TABELA 7). Possivelmente, as seqüências genômicas destes organismos não estavam depositadas nos bancos de dados durante a pesquisa realizada por Dos Santos (2012). Estes novos organismos identificados incluem espécies diazotróficas de referência como *A. brasilense*, *A. lipoferum* e *Azoarcus sp*..

Outros resultados revelados após a análise de vizinhança da proteína DraT, se baseiam nos grupos de estratégia. No grupo de estratégia mais próximo, o de -1 a +1, é possível observar a predominância dos genes *draG*, *nif*A, *nifH* e *nifB*. O grupo de estratégia intermediário, o -3 a +3, possui além dos genes do grupo anterior, mais genes *nif* (*nifD*, *nifK* e *nifX*) além da proteína Ferredoxina III. E no último grupo (-5 a +5), observa-se além dos genes dos grupos anteriores, que existem outros gene mais *nif* (*nifE*, *nifN* e *nifT*) e proteínas ferredoxinas (TABELA 8).

TABELA 8 - RESULTADOS DOS GRUPOS DE ESTRATÉGIA PARA A ANÁLISE DE VIZINHANÇA DO GENE *draT*

CLUSTER	***DE -1 A +1***	***DE -3 A +3***	***DE -5 A +5***
0	*draG*	*draG*	*draG*
1	*nifH*	*nifH*	*nifE* / *nifN* / *nifD*
2	*nifB*	*nifD*	*nifK*
3	*nifA*	*nifK*	*nifH*

4		*nifB*	*nifB*
5		Ferredoxin III	Ferredoxin III
6		*nifX*	*nifX*
7		*nifA*	*nifA*
8			nitrogenase molybdenum-iron protein subunit alpha
9			*nifT*
10			4Fe-4S ferredoxin
12			ferredoxin V

Com o auxilio do software STRING, foi possível observar a interação entre a proteína DraT e seus vizinhos. Nos organismos *Rhodopseudomonas palustris BisB5* e *Geobacter daltonii FRC-32* é possível identificar a DraG (RPD_2408 e Geob_2589, respectivamente) além de protínas Nif e ferredoxinas (TABELA 8), (FIGURA 10).

Os vizinhos da DraT encontrados na análise de vizinhança confirmam a afirmação de que esta proteína é restrita à bactérias diazotróficas. Pois, os genes *nif* *nifH*, *nifB*, *nifD*, *nifN*, *nifK* e *nifE* são os genes básicos para se predizer um organismo como diazotrofo (DOS SANTOS et. al., 2012) e os demais *nifX*, *nifA* e *nifT*, também estão relacionados à nitrogenase (TRIPLETT, 2000). As proteínas Ferredoxinas doam elétrons para a nitrogenase a fim de que a mesma complete seu ciclo catalítico (VALENTINE, 1964; BRUSCHI e GUERLESQUIN, 1988), fato que justifica a proximidade entre os genes analisados.

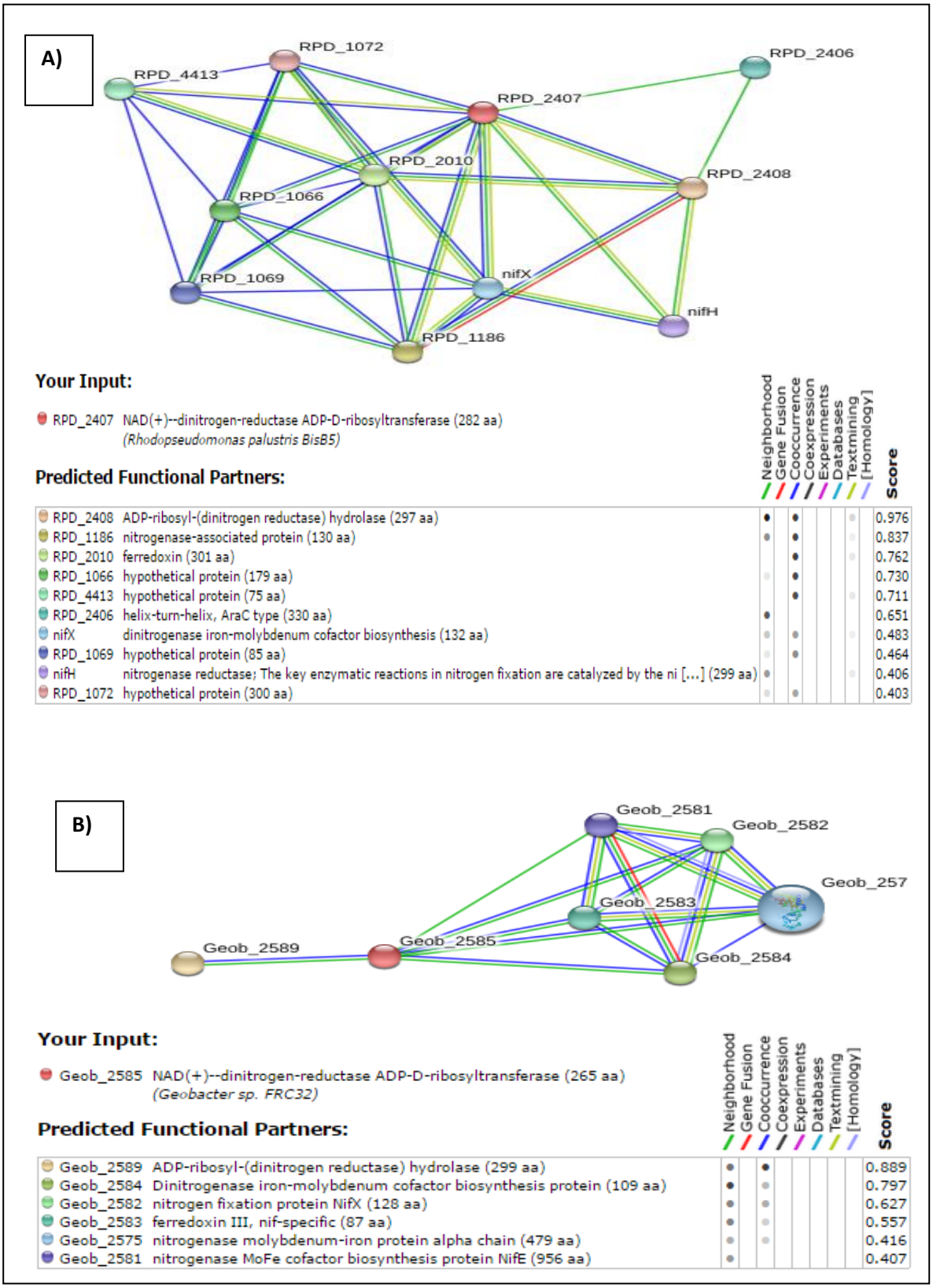

FIGURA 10 – EXEMPLO DE ANÁLISE DE VIZINHANÇA DO GENE *draT* EM (A) *Rhodopseudomonas palustris BisB5: draG* (RPD_2408), *nifH* e *nifX,* ferredoxin (RPD_2010). (B) EM *Geobacter daltonii FRC-32*: *draG* (Geob_2589), ferredoxina III (Geob_2583), *nifD* (Geob_2575), *nifX* (Geob_2582) e *nifE* (Geob_2581)

FONTE: STRING (2014)

A terceira etapa (Agrupamento 3) revelou que proteínas similares à enzima DraG são encontradas em todos os três domínios da vida: *Bacterias*, *Archaea* e *Eukarya*, sendo totalizados 1405 organismos diferentes (FIGURA 11). É possível observar sua predominância no domínio *Bacteria*.

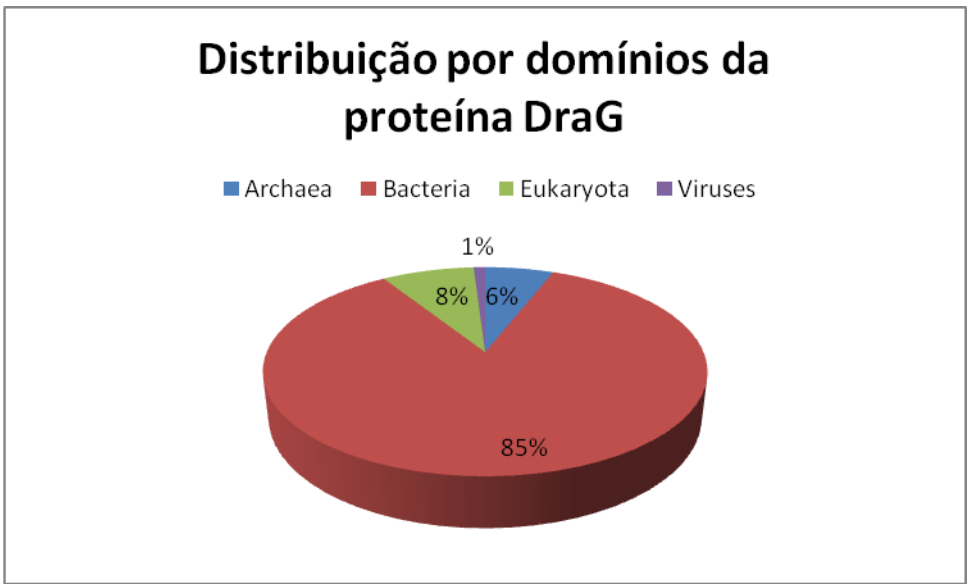

FIGURA 11 – DISTRIBUIÇÃO DA ENZIMA DraG
FONTE: A autora (2014)

A distribuição destes organismos por grupos taxonômicos revela a predominância dos filos *Actinobacteria*, *Firmicutes* e *Proteobacteria* do domíno *Bacteria* (TABELA 9).

TABELA 9 - DISTRIBUIÇÃO POR GRUPOS TAXONÔMICOS DA DraG

GRUPOS TAXONÔMICOS		
DOMÍNIOS	**FILOS**	**QUANTIDADE**
Archaea	*Crenarchaeota*	6
	Euryarchaeota	78
	Thaumarchaeota	1
	TOTAL	**85**
	Actinobacteria	284
	Aquificae	4
	Armatimonadetes	1
	Bacteroidetes	63

Bacteria	*Chlorobi*	4
	Chloroflexi	9
	Chrysiogenetes	1
	Cyanobacteria	49
	Deferribacteres	3
	Deinococcus-Thermus	11
	Firmicutes	266
	Fusobacteria	1
	Gemmatimonadetes	1
	Nitrospinae	1
	Planctomycetes	14
	Proteobacteria	445
	Spirochaetes	10
	Synergistetes	4
	Tenericutes	2
	Thermodesulfobacteria	1
	Thermotogae	4
	Verrucomicrobia	9
	Unclassified	4
	TOTAL	**1191**
Eukaryota	*Alveolata*	2
	Fungi	48
	Metazoa	49
	Viridiplantae	3
	Unclassified	9
	TOTAL	**111**
Viruses	Unclassified	18
	TOTAL	**18**
TOTAL GERAL		**1405**

A etapa baseada na análise de vizinhança do gene *draG* (Agrupamento 4), revelou que independente dos grupos de estratégia (-1 a +1, -3 a +3 e -5 a +5) que totalizam 301 grupos, os genes vizinhos com maior ocorrência são genes cujos produtos codificam para proteínas homólogas a DraG, uma possível riboquinase da subfamília pFKB possivelmente envolvido na fosforilação de açúcares, e um possível transportador de nucleosídeos distruibuídos em 20, 18 e 3 grupos, respectivamente (APÊNDICE 3).

Em *E. coli,* uma proteína similar a DraG é codificada pelo gene *yegU,* mas sua função é ainda desconhecida. Neste organismo, este gene está agrupado no operon *yegTUV* (FIGURA 12). O gene *yegT* codifica para um possível transportador de nucleosideos, *yegU* para proteína similar a DraG e *yegV* para possível riboquinase

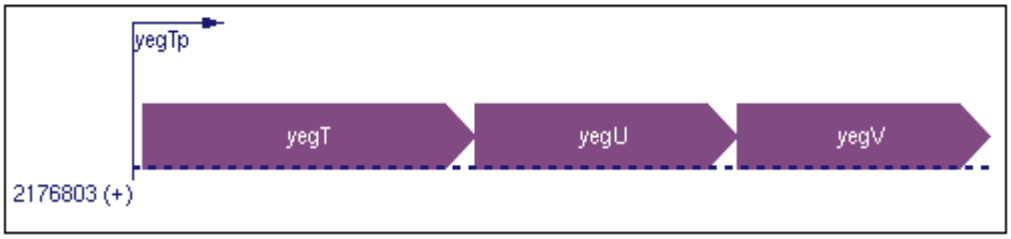

FIGURA 12 – MODELO DO OPERON *yegTUV* DE *Escherichia coli*
FONTE: Ecocyc (2014)

Na análise de vizinhança do gene *draG* observou-se a presença recorrente de genes relacionados ao processo de transporte de nucleosídeos e de uma possível riboquinase (TABELA 10). Esta análise sugere que uma organização gênica semelhante ao operon *yegTUV* de *E. coli* seja conservada em outros organismos.

TABELA 10 - DISTRIBUIÇÃO DOS GENES *yegT, yegU* e *yegV*

Gene	**Anotação do produto**	**-5 a +5 Cluster**	**Taxonomia**
yegT	yegT nucleoside transporter putative nucleoside transport protein	0	Bacteria > Proteobacteria > Gammaproteobacteria

	MFS family transport protein		
yegU	ADP-ribosylglycohydrolase-family protein	8	Bacteria > Firmicutes > Clostridia
		10	Bacteria > Firmicutes > Bacilli
	ADP-ribosylglycohydrolase	12	Bacteria > Proteobacteria > Gammaproteobacteria
		14	Bacteria > Cyanobacteria > Oscillatoriophycideae
	dinitrogenase reductase activating glycohydrolase	17	Bacteria > Cyanobacteria > Gloeobacteria
		73	Bacteria > Spirochaetes > Spirochaetales
			Bacteria > Bacteroidetes > Sphingobacteriia
	ADP-ribosylation/Crystallin J1	144	Bacteria > Cyanobacteria > Nostocales
		155	Bacteria > Actinobacteria > Actinobacteridae
yegV	kinase		Bacteria > Proteobacteria > Gammaproteobacteria
	sugar kinase YegV	2	Bacteria > Actinobacteria > Actinobacteridae
	putative kinase	4	Bacteria > Proteobacteria > Alphaproteobacteria
		6	
	PfkB family kinase	7	Bacteria > Thermobaculum
		13	Bacteria > Firmicutes > Bacilli
	hydroxyethylthiazole kinase		Bacteria > Firmicutes >

	Ribokinase		Clostridia Bacteria > Cyanobacteria > Oscillatoriophycideae

É bem provável que as proteínas homólogas a DraG encontradas durante este trabalho tenham mantido a mesma função biológica durante o processo evolutivo, ou seja, sejam capazes de clivar ligações N-glicosídicas. Esta hipótese é suportada pelo fato de que a enzima DraG de *A. brasilense* e a proteína ARH3 humana apresentam a mesma atividade enzimática (clivagem de ADP ribose N-ligada a arginina) e alta similaridade estrutural.

A análise de vizinhança mostrou uma alta correlação entre homólogos de DraG e proteínas transportadoras de nucleosídeos e uma provável proteína capaz de fosforilar riboses (Riboquinase). Uma hipótese é que estas proteinas atuem em conjunto em uma via de utilização e/ou reciclagem de nuclesídeos e/ou derivados. Por exemplo, a proteína YegT transportaria nucleosídeos, YegU (homológa a DraG) poderia clivar a ligação N-glicosídica liberando a ribose e base nitrogenada. A enzima YegV (riboquinase) poderia então fosforilar a ribose para ser utilizada em outras vias metabólicas (FIGURA 13). Entretanto, esta hipótese requer certamente alguma confirmação experimental.

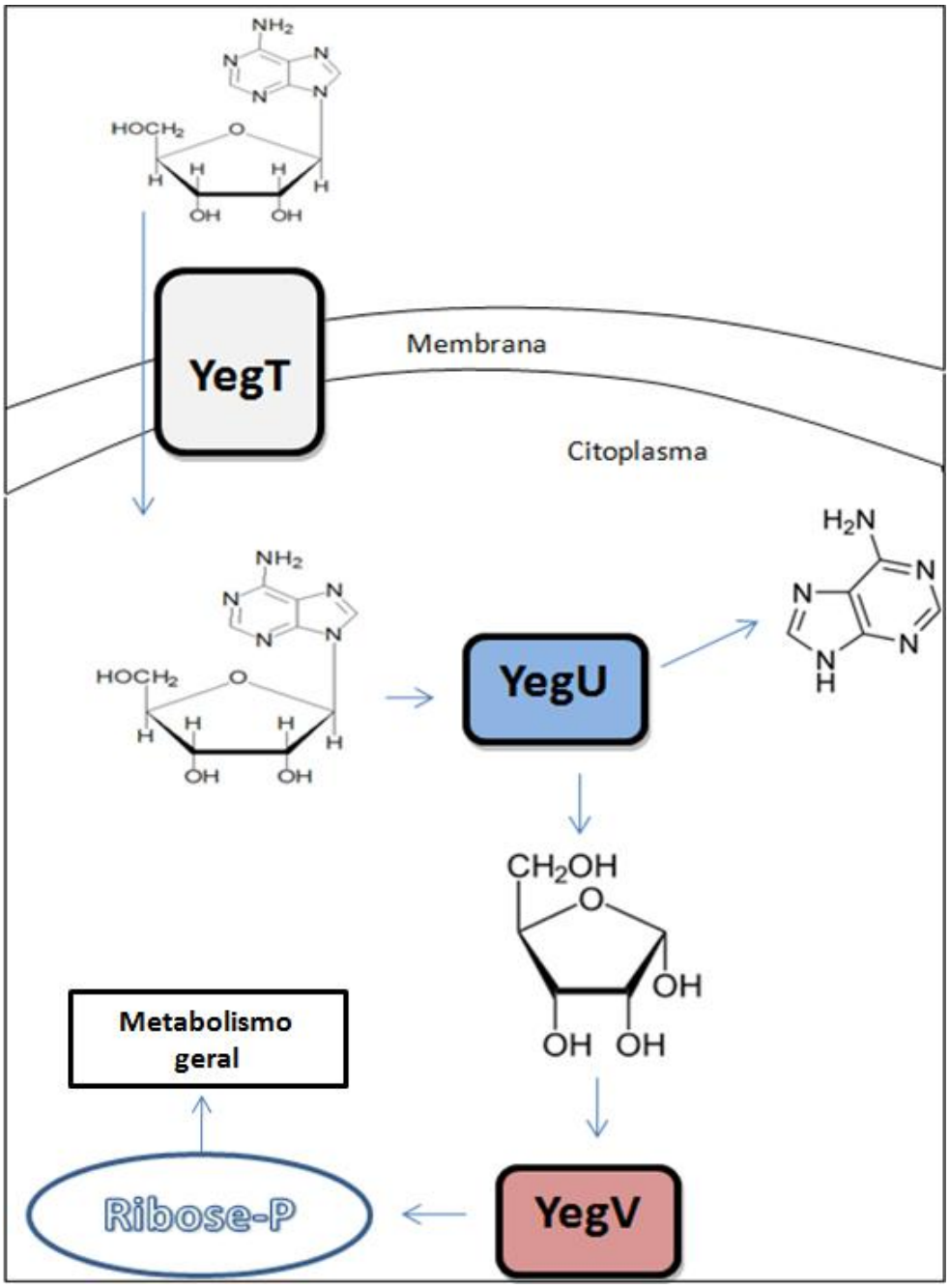

FIGURA 13 – MODELO SUGERIDO
FONTE: A autora (2015)
Nota: Proteína YegT transportando uma adenosina. YegU (homológa a DraG) clivando a ligação N-glicosídica e liberando a ribose e a base nitrogenada. A enzima YegV (riboquinase) fosforilando a ribose.

5 CONCLUSÃO

Neste trabalho desenvolvemos uma análise de bioinformática para identificação de genes homólogos, realização de clusterização destes genes e análise de vizinhança para os genes *draT* e *draG* de *A. brasilense*. Estas análises revelaram que organismos que possuem genes homólogos ao gene *draT* são somente organismos fixadores de nitrogênio. Em todos os casos analisados, o gene *draT* encontra-se próximo ao *cluster* de genes *nif*. Portanto, é muito provável que a única função de proteínas homólogas a DraT seja a catalisar a ADP-ribosilação da nitrogenase.

Por outro lado, genes homólogos ao gene *draG* de *A. brasilense* são comuns incluindo organismos não fixadores de nitrogênio. Esta análise suporta a hipótese que tais proteínas desempenham outra função que não seja somente regular a atividade da nitrogenase. A análise de vizinhança sugere que alguns destes homólogos participem de vias de utilização e/ou reciclagem de nucleosídeos.

O desenvolvimento da ferramenta Blaster colaborou com os resultados encontrados no trabalho de Moure, *et.al.* (2014) e na identificação de 27 possíveis novos organismos fixadores de nitrogênio não descritos em Dos Santos, *et.al.* (2012).

REFERÊNCIAS

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E. W., LIPMAN, D.J. **Basic Local Alignment Search Tool**. J. Mol. Biol. v. 215, p. 403-410, 1990.

BAYAT, A. **Science, medicine, and the future Bioinformatics.** BMJ, v. 324 p. 1018-22, 2002.

BELL, C. E. and EISENBERG, D. **Crystal structure of diphtheria toxin bound to nicotinamide adenine dinucleotide**. Biochemistry. v. 35, p. 1137-1149, 1996.

BENSON, D.A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J., RAPP, B.A., WHEELER, D.L. **GenBank. Nucleic Acids Research**, vol 36 (Database ISSUE), p. D25-D30. 2008.

BERTHOLD, C.L., WANG, H., NORDLUND, S., HÖGBOM, M. **Mechanism of ADP-ribosylation removal revealed by the structure and ligand complexes of the dimanganese mono-ADP-ribosylhydrolase DraG**. Proceedings of the National Academy of Sciences of the United States of America.106, 34. p.14247-14252, 2009.

BLAST, **Blast program selection guide**. Disponível em: < http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=ProgSelectionGuide>. Último acesso em: 20/12/14.

BURRIS, R. H. **Nitrogenases**. J. Biol. Chem. v. 266, p. 9339-9342, 1991.

BRUSCHI, M. e GUERLESQUIN, F. **Structure, function and evolution of bacterial ferredoxins**. FEMS Microbiol. Rev. 4: p. 155-175, 1998.

CANNON, F.C., DIXON, A., POSTAGE, J.R. **Chromosomal Integration of Klebsiella Nitrogen Fixation Genes in Escherichia coli.** Journal of General Microbiology. 80: p. 227-239, 1974.

DANDEKAR, T., SNEL, B., HUYNEN, M., BORK, P. **Conservation of gene order: a fingerprint of proteins that physically interact**. Trends Biochem. Sci. 23: p. 324-328, 1998.

DIXON, R.A.; AUSTIN, S.; BUCK, M.; DRUMMOND, M.; HILL, S.; HOLTEL, A.; MACFARLANE, S.; MERRICK, M.; MINCHIN, S. **Genetics and regulation of nif and related genes in Klebsiella pneumoniae**. Philosophical Transactionns of the Royal Society, London, Série B, v.317, p.147, 1986.

DIXON, R., and KAHN, D. **Genetic regulation of biological nitrogen fixation**. Nat Rev Microbiol 2: p. 621-631, 2004.

DOS SANTOS, P. C., FANG, Z., MASON, S. W., SETUBAL, J.C., DIXON, R. **Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial**. BMC Genomics **13**: p.162, 2012.

EADY, R. R. **Enzymology in free-living diazotrophs**. In: BROUGHTON, W. J.; PUHLER, S. (Ed.) Nitrogen Fixation, v.4, p.1-49, 1986.

ECOCYC. **GENE yegT**. 2013. Disponível em: < http://ecocyc.org/ECOLI/NEW-IMAGE?type=GENE&object=G7130>. Último acesso em: 10/01/15.

EMBL, H., LION B. A. G. **Twilight zone of protein sequence alignments**. Protein Engineering, v.12, n. 2, p. 85-94,1999.

FASSLER, J. e COOPER, P. **BLAST Help Manual, BLAST Glossary**. 2011. Disponível em: < http://www.ncbi.nlm.nih.gov/books/NBK62051/>. Último acesso em: 20/12/14.

HARTMANN, A., FU, H. & BURRIS,R.H. **Regulation of nitrogenase activity by ammonium chloride in Azospirillum spp**. J Bacteriol 165, 864-870,1986.

HOTTIGER, M.O., HASSA, P.O., LÜSCHER, B. **Toward a unified nomenclature for mammalian ADP-ribosyltransferases**. Trends Biochem Sci. v. 35, p. 208–219, 2010.

HSU, H. **Advanced Data Mining Technologies in Bioinformatics**. Idea Group Publishing, 329P, 2006.

HUERGO, L.F., PEDROSA, F.O., MULLER-SANTOS,M., CHUBATSU,L.S., MONTEIRO, R.A., MERRICK, M., and SOUZA, E.M. **PII signal transduction proteins: pivotal players in post-translational control of nitrogenase activity**. Microbiol 158: 176-190, 2012.

HUERGO, L.F., CHUBATSU, L.S., SOUZA, E.M., PEDROSA, F.O., STEFFENS, M.B.R., and MERRIK, M. **Interactions between PII proteins and the nitrogenase regulatory enzymes DraT and DraG in *Azospirillum brasilense***. Febs Lett. *580*, 5232–5236, 2006.

JOUANNEAU, Y.; C. M. MEYER; P. M. VIGNAIS. **Regulation of nitrogenase activity through iron protein interconversion into an active and inactive for in *Rhodopseudomonas capsulata***. Biochim. Biophys. Acta. v. 749, p. 318-328, 1983.

LESK, A. M. **Introdução à Bioinformática**. 2ª Ed. Editora Artmed. p. 101-102, 2008.

LI, W., FU, L., NIU, B., ZHU, Z. **CD-HIT: accelerated for clustering the next-generation sequencing data**. Oxford. Bioinformatics. v. 28, p. 3150-3152, 2012.

LI, X-D, HUERGO, L.F., GASPERINA, A. **Crystal structure of dinitrogenase reductase-activating glycohydrolase (DRAG) reveals conservation in the ADP-ribosylhydrolase fold and specific features in the ADP-ribose-binding pocket**. J Mol Biol v.390, p.737–746, 2009.

LIANG, J. H., NIELSEN, G. M., LIES, D.P. **Mutations in the draT and draG genes of Rhodospirillum rubrum result in loss of regulation of nitrogenase by reversible ADP-ribosylation**. J Bacteriol. v. 173, p.6903–6909, 1991.

LJUNGSTROM, E., YATES, M.G., and NORDLUND, S. **Purification of the activating enzyme for the Fe protein of nitrogenase from *Azospirillum brasilense***. Biochim. Biophys. Acta *994*, p. 210–214, 1989.

LOMAN, N., WATSON, M. **So you want to be a computational biologist?** Nature Biotechnology 31, 11, p. 996–998, 2013.

LOWERY, R.G., and LUDDEN, P.W. **Effect of nucleotides on the activity of dinitrogenase reductase ADP-ribosyltransferase from *Rhodospirillum rubrum***. Biochemistry (Mosc.) *28*, p. 4956–4961, 1989.

LOWERY, R.G., SAARI, L.L., and LUDDEN, P.W. **Reversible Regulation of the Nitrogenase Iron Protein from *Rhodospirillum rubrum* by ADP-Ribosylation *in vitro***. J. Bacteriol. *166*, p. 513–518, 1986.

MICROSOFT. **Novidades no Microsoft Office Excel 2007**. Disponível em: < http://office.microsoft.com/pt-br/excel-help/novidades-no-microsoft-office-excel-2007-HA010073873.aspx>. Último acesso em: 20/10/14.

MOURE, V.R., COSTA, F.F., CRUZ, L.M., PEDROSA, F.O., SOUZA, E.M., LI, X-D, WINKLER, F., HUERGO, L.F. **Regulation of Nitrogenase by Reversible Mono-ADP-Ribosylation**. Current Topics in Microbiology and Immunology, v. 384, p. 89-106, 2014.

NCBI, **The NCBI Handbook, Editores: Jo McEntyre, Jim Ostell. National Center for Biotechonology Information, Bethesda (MD): National Center for Biotechonology Information (US);** 2002. Disponível em: < http://www.ncbi.nlm.nih.gov/books/NBK21101/>. Último acesso em: 01/02/14.

NORDLUND, S. e HÖGBOM, M. **ADP-ribosylation, a mechanism regulating Nitrogenase activity**. FEBS Journal. v. 280, p. 3484-3490, 2013.

PEDROSA, F.O. **Fixação biológica de nitrogênio: fértil idéia**. Ciência hoje, v. 6, p.12-13, 1987.

PELCZAR, M.J., CHAN, E.C.S., KRIEG, N.R. **Microbiologia: conceitos e aplicações**. Volume 2. 2ª Edição. São Paulo: Makron Books, p 517,1996.

POPE, M.R., SAARI, L.L., and LUDDEN, P.W. **N-glycohydrolysis of adenosine diphosphoribosyl arginine linkages by dinitrogenase reductase activating glycohydrolase (activating enzyme) from *Rhodospirillum rubrum*.** J. Biol. Chem. 261, 10104–10111, 1986.

POSTGATE, J. R. **The fundamentals of nitrogen fixation**. Cambridge, Cambridge Univ.Press. 252p, 1982.

PRLIC, A., YATES, A., BLIVEN, S. E., ROSE, P. W., JACOBSEN, J., TROSHIN, P. V., CHAPMAN, M., GAO, J., KOH, C.H., FOISY, S., HOLLAND, R., RIMSA, G., HEUER, M. L., BRANDSTÄTTER-MÜLLER, H., BOURNE, P. E., WILLIS, S. **BioJava: an open-source framework for bioinformatics in 2012**. Bioinformatics. vol 28, p.2693-2695, 2012.

PROSDOCIMI, F. **INTRODUÇÃO A BIOINFORMÁTICA**. 2007.

RAYMOND, J.; SIEFERT, J.L.; STAPLES, C.R.; BLANKENSHIP, R.E. **The natural history of nitrogen fixation**. Mol. Biol. Evol., 21p.541-554, 2004.

RAZA, K. **Application of data mining in bioinformatics**. Indian Journal of Computer Science and Engineering. Vol.1, 2 p.114-118, 2011.

SAARI, L.L., TRIPLETT, E.W., and LUDDEN, P.W. **Purification and Properties of the Activating Enzyme for Iron Protein of Nitrogenase from the Photosynthetic Bacterium *Rhodospirillum rubrum***. J. Biol. Chem. 259, 15502–15508, 1984.

SAARI, L.L., POPE, M., MURRELL, S., and LUDDEN, P. **Studies on the activating enzyme for iron protein of nitrogenase from *Rhodospirillum rubrum***. J. Biol. Chem. 261, 4973, 1986.

SEEFELDT, L.C., HOFFMAN, B.M., and DEAN, D.R. **Mechanism of Mo-Dependent Nitrogenase**. Annu. Rev. Biochem. *78*, 701–722, 2009.

TRIPLETT, E.W. **Prokaryotic nitrogen fixation – a model system for the analysis of a biological process**. Norfolk, UK: Horizon Scientific Press, p. 800, 2000**.**

TSUGE, H., NAGAHAMA, M., NISHIMURA, H., HISATSUNE, J., SAKAGUCHI, Y., ITOGAWA, Y., KATANUMA, N., SAKURAI, J. **Crystal structure and site-directed mutagenesis of enzymatic components from Clostridium perfringens iota-toxin**. J Mol Biol. v.325, p. 471–483, 2003.

UNIPROT, **UniProtKB**. 2002 – 2015 UniProt Consortium. Disponível em: < http://www.uniprot.org/help/uniprotkb>. Último acesso em: 23/01/15.

VALENTINE, R.C. **Bacterial ferredoxin**. Bacteriol Rev. 28: p. 497–517,1964.

VERLI, H. **Bioinformática da Biologia à flexibilidade molecular**. 282 p, 2014.

VISSCHEDYK, D., ROCHON, A., TEMPEL, W., DIMOV, S., PARK,. H. W., MERRILL, A.R. **Certhrax toxin, an anthrax-related ADP-ribosyltransferase from Bacillus cereus**. J Biol Chem. v. 49, p. 41089–41102, 2012.

VON MERING, C., HUYNEN, M., JAEGGI, D., SCHMIDT, S., BORK, P., SNEL, B. **STRING: a database of predicted functional associations between proteins**. Nucleic Acids Res. 1;31(1), p. 258-61, 2003.

ZHANG, Y.; BURRIS, R. H.; LUDDEN, P. W.; ROBERTS, G. P. **Regulation of nitrogen fixation in *Azospirillum brasilense***. FEMS Microbiol. Rev. v. 152, p.195-204, 1997.

ZUMFT, W. G. & CASTILHO, F. **Regulatory properties of the nitrogenase from *Rhodopseudomonas palustris***. Arch Microbiol 117, 53-60,1978.

APÊNDICES

APÊNDICE 1 – SEQUÊNCIAS DE AMINOÁCIDOS UTILIZADAS NO BLASTP.

>gi|392381758|ref|YP_005030955.1| ADP-ribosyl-[dinitrogenase reductase]transferase [Azospirillum brasilense Sp245]
MADGSASGGLRDARGRLIDPDSPLALKAHRTNLLNVTAEALCAESFNDEPRRLRIGGTRSEH
AVLFEALDESPNSLIAADIFQHYMAFTFGLNPDFSGAEGVDGRRRYRASYLRLLKGWLFDSN
NAEGAVLKGWVESRFGLLPTYHKEPILRFASDAWRTYGEEKVATRFHNNNINLQLDLLYEYC
QWWMRRGWPDAPMPSRASHIRLYRGVNDFEEHHIVARPDKRTAVLRLNNLSSFSIERDIAGQ
FGDYILETWVPLTKVVFFRDILPRYPFKGEGEYLVVGGDYRVGVSLL

>gi|392381759|ref|YP_005030956.1| ADP-ribosyl-[dinitrogenase reductase]hydrolase [Azospirillum brasilense Sp245]
MTDHSIRSRALGAYLGLACGDALGATVEFLTKGEIAHQYGVHKHIKGGGWLKLPAGQVTDDT
EMSIHLGRAILAAPEWDARRAAEEFAVWLKGVPVDVGDTTRRGIRRFIMHGTLSEPESEYHA
GNGAAMRNLPVALATLGDDDAFERWTVEQAHITHCNAMSDAATLTLGHMVRRLVLGGSVRDV
RDESNKLIAQHRQFKFQPYRGLATAYIVDTMQTVMHYYFQTDSVESCVVETVNQGGDADTTG
AIAGMLAGATYGVETIPPRWLRKLDRDVYNEICAQVDGLLARSPALKQG

>gi|392381754|ref|YP_005030951.1| nitrogenase iron protein, nifH; dinitrogenase reductase [Azospirillum brasilense Sp245]
MSLRQIAFYGKGGIGKSTTSQNTLAALVELDQKILIVGCDPKADSTRLILHAKAQDTVLHLA
AEAGSVEDLELEDVLKIGYKGIKCVESGGPEPGVGCAGRGVITSINFLEENGAYDDVDYVSY
DVLGDVVCGGFAMPIRENKAQEIYIVMSGEMMALYAANNIAKGILKYAHSGGVRLGGLICNE
RQTDKEIDLASALAARLGTQLIHFVPRDNIVQHAELRRMTVIEYAPDSQQAQEYRQLANKVH
ANKGKGTIPTPITMEELEEMLMDFGIMKSEEQQLAELQAKEAAKA

>gi|392381753|ref|YP_005030950.1| nitrogenase molybdenum-iron protein alpha chain,nifD [Azospirillum brasilense Sp245]
MSLSVNEGVDVKGLVDKVLEAYPEKSRKRRAKHLNVLEAEAKDCGVKSNIKSIPGVMTIRGC
AYAGSKGVVWGPIKDMIHISHGPVGCGYYSWSGRRNYYVGDTGVDSWGTMHFTSDFQEKDIV
FGGDKKLHKVIEEINELFPLVNGISIQSECPIGLIGDDIEAVARAKSEELGKPVVPVRCEGF
RGVSQSLGHHIANDVIRDWIFEKTEPKEGFVSTPYDVTIIGDYNIGGDAWASRILLEEIGLR
VIAQWSGDGTLAELENTPKAKVNLIHCYRSMNYIARHMEEKFGIPWMEYNFFGPSQIAESLR
KIAALFDDTIKENAEKVIAKYQPMVDAVIAKFKPRLEGKKVMIYVGGLRPRHVVDAYHDLGM
EIVGTGYEFAHNDDYQRTQHYVKEGTLIYDDVTAFELEKFVEVMRPDLVASGIKEKYVFQKM
GLPFRQMHSWDYSGPYHGYDGFAIFARDMDLAINNPVWGIMKAPF

>gi|392381748|ref|YP_005030945.1| nitrogenase molybdenum-cofactor synthesis protein nifE [Azospirillum brasilense Sp245]

MLQEKLQDVFNEPGCSTNQAKSEKERKKGCSKALKPGAAAGGCAYDGAMIALQPIADAAHLV
HGPIACLGNSWDNRGTKSSGSQLYRTGFTTDMSELDIIHGGEKKLYKAIKEIVQQYDPPAVF
VYQTCVPAMIGDDIEAVCKFAAKKLGKPVIPVMAPGFVGSKNLGNKLAGETLLDHVIGTVEP
EVTTPTDICIVGEYNLAGELWLVKPLLDEIGIRILSCISGDGRYNEVAQAHRARLTMVVCSQ
ALVNVGRKMQERWGIPYFEGSFYGVSDMSDTLRTMARMLVERGADKAIIDRTEGVIAREESR
VWRRLEPYKPRFDGKRVLLFTGGVKSWSMVTALEGAGLTIVGTSTKKSTKEDKERIKKMKGE
EFHQWDDLKPRDIYKMLRDSEADIMMSGGRSQFIALKAKVPWLDLNQERHTPYAGYDGIVNL
CEEIDKTLSNPIWRQVRLAAPWDLKPDAKPDARPVGA

>gi|392381752|ref|YP_005030949.1| nitrogenase molybdenum-iron protein beta chain,nifK [Azospirillum brasilense Sp245]
MSHPVSQSADKVIDHFTLFRQPEYKELFERKKTEFEYGHSDEEVARVSAWTKTEEYKEKNFA
REAVVINPTKACQPIGAMFAAQGFEGTLPFVHGSQGCVAYYRTHLTRHFKEPNSAVSSSMTE
DAAVFGGLNNMIDGLANAYALYKPKMIAVLTTCMAEVIGDDLSGFINNAKNKESVPADFPVP
FAHTPAFVGSHIVGYDNMIKGVLTHFWGTSENFDTPKNEKINLIPGFDGFAVGNNRELKRIA
GLFGIDLTILSDVSDNFDTPADGEYRMYDGGTPLEATKEAVHAKATISMQEYCTPQSLQFIK
EKGQQVAKYNYPMGVTGTDELLLKLAELSGKPVPAELKLERGRLVDAIADSHTHLHGKRFAV
YGDPDFCLGMSKFLMELGAEPVHILSTSGSKKWEKQVQKVLDASPFGKSGKAYGGKDLWHLR
SLLFTDKVDYIIGNSYGKYLERDTKIPLIRLTYPIFDRHHHHRYPTWGYQGALNVLVRILDR
IFEDMDANTNIVGETDYSFDLVR

>gi|392381747|ref|YP_005030944.1| nitrogenase reductase-associated ferredoxin,nifN [Azospirillum brasilense Sp245]
MGNIQRFPHSAKAASTNPLKMSQPLGAALAFLGVDRCMPLFHGSQGCTAFGLVLLVRHFREA
IPLQTTAMDQVSTILGGYENLEQAVRTIHERNAPALIGVATTGVTETKGEDMSGQYSLFRQR
NPALAGLKLVFANTPDFSGGFEDGFSAAVTGIVEEVVQPSETTMKGQINVLAGCHLSPGDVE
ELRDIIESFGLSPIFLPDLSLSMSGRQPDDFTATSLGGVTVDQIASMGASELTLVIGEHMRV
AAAALELKTDVRSVFFDRLTGLEASDRLVRTLSELSGRPVPAKLRRQRETLVDGMLDGHFFY
SRKRIAVALEPDLLYAVTSFLADMGAEVIAAVSPTQTAVLEKLKAATVMVGDHSDVETLARD
ADLIVSNSHGRQGAARIGVPLHRMGLPMFDRLGAGLKVHVGYRGTRELLFEIGNLFLSREMD
HDEHGHAHGHRHGDGHEHGQHCGSGSCGCSAG

>gi|392381771|ref|YP_005030968.1| nitrogenase FeMo cofactor biosynthesis protein NifB [Azospirillum brasilense Sp245]
MANVISLDSILGVGELKAAAEAPPAAASGCASSSCGSSDGPADMAPEVWEKVKNHPCYSEEA
HHYFARMHVAVAPACNIQCNYCNRKYDCSNESRPGVVSEKLTPDQALRKIMAVAKEIPQLSV
IGIAGPGDSLAAGGKNTFKTFEMLAKKAPDLKLCLSTNGLALPDHVDTIANYNIDHVTITIN
MVDPEVGQHIYPWIFHDHKRWTGLDAAKILHERQMLGLEMLTSRGILVKVNSVMIPGINDEH
LMDVNKAVKSRGAFLHNIMPLISDPAHGTHFGLTGQRGPTAQELKVLQDQCEGGAKLMRHCR
QCRADAVGLLGEDRGSEFTIDQIEAMGEVEYDQEAARTYREHVEGERAGRHAAKAAAQADVA
ETVGTEVQPILIAVATKGGERINEHFGHAKEFQIYEVGPKGAKFVGHRRVDQYCEGGSGDED
ALGGVLSAINDCTAVFVAKIGGCPSTSLKDAGIEPVDRFAFEYIEESALTYFKDYAERLGQG
AINAREGQDAVIRTGAFTALRA

APÊNDICE 2 – *LOCUS TAG* DOS GENES *draT, draG*, *nifH*, *nifE*, *nifD*, *nifN*, *nifB* E *nifK***.**

Organismos	**Gene (*locus tag*)**							
	draT	***draG***	***nifH***	***nifB***	***nifD***	***nifN***	***nifK***	***nifE***
Acidithiobacillus ferrivorans SS3	Acife_1652	Acife_1653	Acife_1650	Acife_1655	Acife_1649	Acife_1644	Acife_1648	Acife_1645
Acidithiobacillus ferrooxidans ATCC 53993	Lferr_1242	Lferr_1243	Lferr_1240	Lferr_1245	Lferr_1239	Lferr_1234	Lferr_1238	Lferr_1235
Allochromatium vinosum DSM 180	Alvin_1882	Alvin_2299	Alvin_1881	Alvin_3039	Alvin_1880	Alvin_0904	Alvin_1879	Alvin_0903
Anaeromyxobacter sp. Fw109-5	Anae109 _ 3033	Anae109 _ 3034	Anae109 _3030	Anae109 _ 3024	Anae109 _ 3029	Anae109 _ 3027	Anae109 _ 3028	Anae109 _3028
Anaeromyxobacter sp. K	AnaeK _3210	AnaeK _0460	AnaeK _3203	AnaeK _3202	AnaeK _3204	AnaeK _3206	AnaeK _3205	AnaeK _3206
Azoarcus sp. BH72	azo0537	azo2644	azo0538	azo0523	azo0539	azo0561	azo0540	azo0562
Azoarcus sp. KH32C	AZKH _0793	AZKH_0794	AZKH_0791	AZKH_4060 / AZKH_0798	AZKH_4079 / AZKH_0790	AZKH _0782	AZKH_4080 / AZKH_0789	AZKH_4081 / AZKH_0783
Azospirillum brasilense Sp245	AZOBR _70132	AZOBR _70133	AZOBR _70128	AZOBR _70147	AZOBR _70127	AZOBR _70121	AZOBR _70126	AZOBR _70122
Azospirillum lipoferum 4B	AZOLI _0684	AZOLI _0685	AZOLI _0681	AZOLI _2311	AZOLI _0680	AZOLI _0673	AZOLI _0679	AZOLI _0674
Azospirillum sp. B510	AZL _007730	AZL _007740	AZL _007710	AZL _022470	AZL _007700	AZL _007650	AZ L_007690	AZL _007660
Calditerrivibrio nitroreducens DSM 19672	Calni_0683	Calni_1418 / Calni_0682	Calni_0684	Calni_0673	Calni_0685	Calni_0679	Calni_0686	Calni_0678
Candidatus Accumulibacter phosphatis clade IIA str	CAP2UW1 _4446	CAP2UW1 _4428	CAP2UW1 _4447	CAP2UW1 _4422	CAP2UW1 _4448	CAP2UW1 _4449	CAP2UW1 _4372	CAP2UW1 _4373
Coraliomargarita akajimensis DSM 45221	Caka_2843	Caka_1534	Caka_2839	Caka_2834	Caka_2838	Caka_2835	Caka_2837	Caka_2836
Dechloromonas aromatica RCB	Daro_1416	Daro_1447	Daro_1415	Daro_1454	Daro_1414	Daro_1505	Daro_1413	Daro_1501
Dechlorosoma suillum PS	Dsui_0954	Dsui_0953	Dsui_0955	Dsui_0943	Dsui_0956	Dsui_0551	Dsui_0957	Dsui_0550
Denitrovibrio acetiphilus DSM 12809	Dacet _1042	Dacet _1043	Dacet _1046	Dacet _1055	Não encontrado	Dacet _1051	Dacet _1048	Dacet_1047 / Dacet_1050
Desulfobacterium autotrophicum HRM2	HRM2 _09830	HRM2 _48570	HRM2 _09670	HRM2 _09740	HRM2 _09700	HRM2 _09730	HRM2 _09710	HRM2 _09720
Desulfurispirillum indicum S5	Selin_0259	Selin_0258	Selin_0690	Selin_2076	Selin_0691	Selin_0263	Selin_0692	Selin_0262
Geobacter bemidjiensis Bem	Gbem _2081	Gbem _2034	Gbem _2074	Gbem _2083	Gbem _2075	Gbem _2077	Gbem _2076	Gbem _2077
Geobacter daltonii	Geob	Geob	Geob	Geob	Geob	Geob	Geob	Geob

FRC-32	_2585	_2589	_2574	_2468	_2575	_2582	_2577	_2581
Geobacter lovleyi SZ	Glov_0630	Glov_0420	Glov_0650	Glov_0422	Glov_0649	Glov_0639	Glov_0648	Glov_0639
Geobacter metallireducens GS-15	Gmet _0674	Gmet _3377	Gmet _0662	Gmet _0681	Gmet _0663	Gmet _0669	Gmet _0664	Gmet _0669
Geobacter sp. M18	GM18 _1869	GM18 _2171	GM18_1862 / GM18_1685	GM18 _1868	Não encontrado	GM18 _1865	GM18 _1864	GM18 _1863
Geobacter sp. M21	GM21 _2137	GM21 _2136	GM21 _2145	GM21 _2138	Não encontrado	GM21 _2142	GM21 _2143	GM21 _2144
Geobacter sulfurreducens PCA	GSU2802	GSU0194	GSU2821	GSU2799	GSU2820	GSU2806	GSU2819	GSU2806
Geobacter uraniireducens Rf4	Gura_1205	Gura_0904	Gura_1175	Gura_1209	Gura_1176	Gura_1201	Gura_1177	Gura_1201
Magnetococcus marinus MC-1	Mmc1 _1203	Mmc1 _1210	Mmc1 _1202	Mmc1 _1206	Mmc1 _1201	Mmc1 _1194	Mmc1 _1200	Mmc1 _1195
Magnetospirillum gryphiswaldense MSR-1	MGMSR _0346	MGMSR _0347	MGMSR _0345	MGMSR _0356	MGMSR _0344	MGMSR _0307	MGMSR _0343	MGMSR _0306
Magnetospirillum magneticum AMB-1	amb1575	amb1576	amb1574	amb1583	amb1573	amb1568	amb1572	amb1569
Methylococcus capsulatus str. Bath	MCA1828	Não encontrado	MCA0229	MCA0204	MCA0230	MCA0234	MCA0231	MCA0233
Methylomonas methanica MC09	Metme _3770	Metme _3769	Metme _1460	Metme _1621	Metme _1461	Metme _3081	Metme _1462	Metme _3080
Pelobacter carbinolicus DSM 2380	Pcar_2092	Pcar_2093	Pcar_2098	Pcar_2107	Pcar_2099	Pcar_0314	Pcar_2100	Pcar_0313
Pelobacter propionicus DSM 2379	Ppro_3487	Ppro_3462	Ppro_3467	Ppro_3463	Ppro_3468	Ppro_3491	Ppro_3469	Ppro_3491
Rhodobacter capsulatus SB 1003	RCAP_rcc 03016	RCAP_rcc0 3017	RCAP_rcc0 0572	RCAP_rcc0 0566	RCAP_rcc00 571	RCAP_rcc0 3279	RCAP_rcc0 0570	RCAP_rcc0 3280
Rhodobacter sphaeroides ATCC 17025	Rsph17025 _3203	Rsph17025 _3202	Rsph17025 _0774	Rsph17025 _1241	Rsph17025 _1247	Rsph17025 _1250	Rsph17025 _1248	Rsph17025 _1249
Rhodopseudomonas palustris BisA53	RPE_0837	RPE_4503	RPE_4533	RPE_4541	RPE_4532	RPE_4529	RPE_4531	RPE_4530
Rhodopseudomonas palustris BisB18	RPC_0878	RPC_0422	RPC_4463	RPC_4472	RPC_4462	RPC_4459	RPC_4461	RPC_4460
Rhodopseudomonas palustris BisB5	RPD_2407	RPD_2408	RPD_1073	RPD_1063	RPD_1074	RPD_1077	RPD_1075	RPD_1076
Rhodopseudomonas palustris CGA009	RPA1431	RPA2406	RPA4620 / RPA1376 / RPA1438	RPA4630	RPA4619 / RPA1378 / RPA1437	RPA4616 / RPA1372	RPA4618 / RPA1380	RPA4617 / RPA1373
Rhodopseudomonas palustris DX-1	Rpdx1 _3108	Rpdx1 _3107	Rpdx1 _4798	Rpdx1 _4808	Rpdx1 _4797	Rpdx1 _4794	Rpdx1 _4796	Rpdx1 _4795
Rhodopseudomonas palustris HaA2	RPB_3044	RPB_3043	RPB_0969	RPB_0959	RPB_0970	RPB_0973	RPB_0971	RPB_0972
Rhodopseudomonas palustris TIE-1	Rpal_1616	Rpal_2658	Rpal_5101	Rpal_5111	Rpal_5100	Rpal_5097	Rpal_5099	Rpal_5098
Rhodospirillum photometricum DSM 122	RSPPHO _02585 / RSPPHO	RSPPHO _02586	RSPPHO _02674	RSPPHO _02535	RSPPHO _02583	RSPPHO _02536	RSPPHO _02582	RSPPHO _02537

	_00485							
Rhodospirillum rubrum ATCC 11170	Rru_A1009	Rru_A1008	Rru_A1010 / Rru_A1395	Rru_A0994 / Rru_A0796	Rru_A1011 / Rru_A1394	Rru_A2285	Rru_A1012 / Rru_A1392	Rru_A2286
Rhodospirillum rubrum F11	F11_05200	F11_05195	F11_04080	F11_05125	F11_05210	F11_11750	F11_05215	F11_11755
Rubrivivax gelatinosus IL144	RGE _42860	RGE _42850	RGE _42880	RGE _42670	RGE _42890	RGE _42960	RGE _42900	RGE _42950
Sideroxydans lithotrophicus ES-1	Slit_0880	Slit_0832	Slit_0881	Slit_0837	Slit_0882	Slit_0904	Slit_0883	Slit_0903
Teredinibacter turnerae T7901	TERTU _0343	TERTU _0344	TERTU _1537	TERTU _1520	TERTU _1538	TERTU _1578	TERTU _1539	TERTU _1577
Thiocystis violascens DSM 198	Thivi_3646	Thivi_2026	Thivi_3647	Thivi_2127	Thivi_3648	Thivi_2739	Thivi_3649	Thivi_2740
Thioflavicoccus mobilis 8321	Thimo _3246	Thimo _2694	Thimo _3245	Thimo _3248	Thimo _3244	Thimo _3646	Thimo _3243	Thimo _3647
Tolumonas auensis DSM 9187	Tola_0578	Tola_0577	Tola_0650	Tola_0604	Tola_0651	Tola_0656	Tola_0652	Tola_0655

APÊNDICE 3 – PRODUTOS DOS GRUPOS GERADOS PELO SOFTWARE CD-HIT REFERENTE A PROTEÍNA DraG

CLUSTER	DE -1 A +1	DE -3 A +3	DE -5 A +5
0	nucleoside transporter / YegT	hydrophilic substrate transporter; MFS superfamily / nucleoside transporter / YegT	hydrophilic substrate transporter; MFS superfamily / nucleoside transporter / YegT
1	putative kinase / PfkB family / yegV	DNA-binding transcriptional regulator/ GntR family transcriptional regulator / putative HTH-type transcriptional regulator YegW	DNA-binding transcriptional regulator/ GntR family transcriptional regulator / putative HTH-type transcriptional regulator YegW
2	cytosine/purines uracil thiamine allantoin permease	kinase / PfkB family / yegV	kinase / PfkB family / yegV
3	PfkB domain-containing protein	fructose-bisphosphate aldolase class 1 / deoxyribose-phosphate aldolase/phospho-2-dehydro-3-deoxyheptonate aldolase	DhnA-type fructose-1,6-bisphosphate aldolase-like enzyme / fructose-bisphosphate aldolase
4	ADP-ribosylation/Crystallin J1 / ADP-ribosylglycohydrolase-like protein	ADP-ribosylation/Crystallin J1	hydroxyethylthiazole kinase
5	DNA hydrolase / NUDIX hydrolase	putative membrane protein / ADP-ribosylation/Crystallin J1	L-carnitine dehydratase/bile acid-inducible protein F / formyl-CoA transferase
6	putative GntR family regulatory protein	ADP-ribosylation/Crystallin J1	PfkB domain-containing protein / ribokinase
7	ADP-ribosylation/Crystallin J1 / ADP-ribosylglycohydrolase-like protein	PfkB domain-containing protein / Ribokinase	bifunctional hydroxy-methylpyrimidine kinase/ hydroxy-phosphomethylpyrimidine kinase / phosphomethylpyrimidine kinase
8	ADP-ribosylation/Crystallin J1 / putative membrane protein	kinase / PfkB family / yegV	ADP-ribosylation/Crystallin J1
9	ADP-ribosylation/Crystallin J1 / ADP-ribosylglycohydrolase	ADP-ribosylglycohydrolase-family protein / ADP-ribosylation/Crystallin J1	putative membrane protein
10	-	-	ADP-ribosylation/Crystallin J1
11	aminoimidazole riboside kinase / carbohydrate kinase / ribokinase	acyl-CoA transferase/carnitine dehydratase / acyl-CoA transferase / formyl-CoA transferase	glutamate synthase (NADPH) large subunit / Ferredoxin-dependent glutamate synthase 1
12	ADP-ribosylglycohydrolase	putative membrane-bound hydrolase / hydrolase / putative glycosyl hydrolase	ADP-ribosylglycohydrolase-family protein / ADP-ribosylation/Crystallin J1
13	-	ADP-ribosylglycohydrolase-family protein / ADP-ribosylation/Crystallin J1	PfkB domain-containing protein / ribokinase YegV
14	-	bifunctional hydroxy-methylpyrimidine kinase/ hydroxy-phosphomethylpyrimidine kinase / phosphomethylpyrimidine kinase	ADP-ribosylglycohydrolase-family protein / ADP-ribosylation/Crystallin J1
15	-	cytosine/purines uracil thiamine allantoin permease / purine-cytosine permease-like transporter	glutamate synthase NADH/NADPH small subunit / glutamate synthase subunit beta
16	-	tagatose-bisphosphate aldolase / D-tagatose 1,6-bisphosphate aldolase 2, catalytic subunit	putative family 25 glycosyl hydrolase / glycosyl hydrolase family protein
17	-	GntR family transcriptional regulator	ADP-ribosylglycohydrolase
18	-	putative hydrolase / NUDIX hydrolase / ADP-ribose pyrophosphatase	-
19	hypothetical protein	glutamate synthase (ferredoxin) / Ferredoxin-dependent glutamate synthase 1 / glutamate synthase(NADPH) large subunit	purine-cytosine permease-like transporter / cytosine/purines uracil thiamine allantoin permease
20	7TM receptor with intracellular metal dependent phosphohydrolase	lipid kinase	tagatose-bisphosphate aldolase / D-tagatose 1,6-bisphosphate aldolase 2, catalytic subunit

21	3-dehydroquinate dehydratase	binding-protein-dependent transport system inner membrane protein	putative hydrolase / NUDIX hydrolase / ADP-ribose pyrophosphatase
22	-	-	GntR family transcriptional regulator
23	-	-	PTS system galactitol-specific transporter subunit IIA
24	hypothetical protein	DNA hydrolase / NUDIX hydrolase	-
25	-	-	citrate lyase subunit beta / citrate (Pro-3S)-lyase, beta subunit
26	binding-protein-dependent transport systems inner membrane component	-	DNA hydrolase / NUDIX hydrolase
27	-	RNA:NAD 2'-phosphotransferase / phosphotransferase KptA/Tpt1	Ribonucleotide reductase of class II (coenzyme B12-dependent) / vitamin B12-dependent ribonucleotide reductase
28	-	-	ABC transporter ATP-binding protein / ABC transporter ATPase
29	Mg2+ transporter MgtE	-	binding-protein-dependent transport systems inner membrane component / sugar ABC transporter permease
31	-	-	lipid kinase
32	hypothetical protein	hypothetical protein	-
33	-	hypothetical protein	-
34	GntR family transcriptional regulator / transcriptional regulator	7TM receptor with intracellular metal dependent phosphohydrolase	PilH protein
35	PfkB domain-containing protein / Ribokinase	-	-
36	-	-	two component transcriptional regulator, winged helix family
37	hypothetical protein	-	-
39	-	Rhs element Vgr protein	phosphotransferase KptA/Tpt1 / RNA 2-phosphotransferase
41	phosphopyruvate hydratase / enolase	family 1 extracellular solute-binding protein / sugar ABC transporter periplasmic protein	-
42	DNA-3-methyladenine glycosylase I	-	short-chain dehydrogenase/reductase SDR
43	hypothetical protein	hypothetical protein	-
44	ABC transporter substrate-binding protein / sugar ABC transporter periplasmic protein	-	ImpA family type VI secretion-associated protein / Rhs element Vgr protein
45	ADP-ribosylglycohydrolase, partial	3-dehydroquinate dehydratase	-
46	-	metallophosphoesterase / diadenosine tetraphosphatase and related serine/threonine protein phosphatase	family 1 extracellular solute-binding protein
48	-	glutamate synthase, NADH/NADPH, small subunit	-
49	-	Nicotinamide-nucleotide adenylyltransferase, NadR family or Ribosylnicotinamide kinase / cytidyltransferase	-
51	hypothetical protein	hypothetical protein	regulatory protein
52	-	-	type IV pilus response regulator PilG / pilus protein
53	family 1 extracellular solute-binding protein / sugar ABC transporter periplasmic protein	nicotinamide mononucleotide transporter PnuC / Ribosyl nicotinamide transporter, PnuC	7TM receptor with intracellular metal dependent phosphohydrolase
54	hypothetical protein	hypothetical protein	FO synthase / 7,8-didemethyl-8-hydroxy-5-deazariboflavin synthase subunit 1 / 7,8-didemethyl-8-hydroxy-5-deazariboflavin synthase subunit 2
55	response regulator receiver / two component transcriptional regulator	-	-
56	metallophosphoesterase / bis(5'-nucleosyl)-	purine-cytosine permease or related protein / cytosine/purines uracil	-

	tetraphosphatase	thiamine allantoin permease	
57	-	[Fe] hydrogenase, electron-transfer subunit / NADH dehydrogenase (ubiquinone) 24 kDa subunit / NAD-reducing hydrogenase subunit HoxE	ABC transporter ATP-binding protein
58	hypothetical protein	NAD-dependent deacetylase 2 / SIR2 family transcriptional regulator	diadenosine tetraphosphatase and related serine/threonine protein phosphatase
59	-	-	hypothetical protein
60	-	-	binding-protein-dependent transporters inner membrane component
61	SARP family transcriptional regulator	NAD-reducing hydrogenase subunit HoxF /NADH dehydrogenase subunit F / [Fe] hydrogenase, electron-transfer subunit	hypothetical protein
62	GntR family transcriptional regulator / HTH-type transcriptional regulator frlR	hypothetical protein	DNA-binding protein
63	-	2-hydroxycyclohexanecarboxyl-CoA dehydrogenase / short-chain dehydrogenase/reductase SDR	hypothetical protein
64	Rhs element Vgr protein / ImpA family type VI secretion-associated protein	sulfite reductase (ferredoxin)	-
65	-	DNA topoisomerase I	-
66	-	Citrate (pro-3S)-lyase subunit beta / citrate lyase subunit beta	-
68	-	-	3-dehydroquinate dehydratase
69	Putative ribokinase / PfkB domain-containing protein	hypothetical protein	chorismate mutase II
70	-	-	hypothetical protein
71	-	-	nuclease-like protein
72	ADP-ribose pyrophosphatase / NUDIX hydrolase / NTP pyrophosphohydrolase	ADP-ribosylation/Crystallin J1 / ADP-ribosylglycohydrolase	-
73	-	sugar ABC transporter permease / binding-protein-dependent transport systems inner membrane component	ADP-ribosylglycohydrolase / ADP-ribosylation/crystallin J1
74	-	Threonyl tRNA synthetase	aldo/keto reductase
75	heme exporter, protein B / cytochrome c-type biogenesis protein CcmB	short-chain dehydrogenase/reductase SDR / glucose 1-dehydrogenase	binding-protein-dependent transporters inner membrane component / ABC-type polysaccharide transport system, permease component
76	-	Mg(2+) transporter mgtE / magnesium transporter	-
79	threonyl-tRNA synthetase	GntR family transcriptional regulator / transcriptional regulator	-
80	-	-	NAD-dependent protein deacetylase of SIR2 family
81	-	binding-protein-dependent transport systems inner membrane component	binding-protein-dependent transport systems inner membrane component / sugar ABC transporter permease
83	-	family 1 extracellular solute-binding protein / ABC transporter substrate-binding protein	-
84	-	ATPase of ABC transporter with duplicated ATPase domains / ABC transporter ATPase	Ribosyl nicotinamide transporter, PnuC / nicotinamide mononucleotide transporter
85	-	fructose 1,6-bisphosphatase II	hypothetical protein
86	-	-	ATP-cone domain-containing protein / transcriptional regulator NrdR
87	-	binding-protein-dependent transport systems inner membrane component / ABC-type polysaccharide transport system, permease component	hypothetical protein
88	-	-	DNA topoisomerase I

89	-	hypothetical protein	-
90	-	transketolase	purine-cytosine permease
91	-	ABC transporter ATP-binding protein / spermidine/putrescine ABC transporter ATPase	D-tagatose 1,6-bisphosphate aldolase 2 subunit
92	-		ABC transporter
93	-	-	cytidyltransferase / Nicotinamide-nucleotide adenylyltransferase, NadR family or Ribosylnicotinamide kinase
94	-	-	binding-protein-dependent transporter inner membrane component
96	-	cell wall endopeptidase / peptidase m23b	NAD-reducing hydrogenase subunit HoxE / [Fe] hydrogenase, electron-transfer subunit
98	-	5'(3')-nucleotidase/polyphosphatase / stationary phase survival protein SurE	-
100	-	histidine triad (HIT) protein	nickel/cobalt homeostasis protein RcnB
101	-	hypothetical protein	-
102	-	-	sulfite reductase (ferredoxin)
103	-	-	[Fe] hydrogenase, electron-transfer subunit / NADH dehydrogenase subunit F
107	-	response regulator receiver / two component transcriptional regulator, winged helix family	-
108	-	phosphoprotein phosphatase / metallophosphoesterase	-
109	-	-	transcriptional repressor cytR / LacI family transcriptional regulator
110	-	putative cyclase/dehydrase	monooxygenase / luciferase family oxidoreductase
111	-	-	CDP-diacylglycerol/serine O-phosphatidyltransferase / phosphatidylserine synthase
112	-	hypothetical protein	-
113	-	putative nicotinate phosphoribosyltransferase / nicotinate phosphoribosyltransferase	putative ribitol 2-dehydrogenase
114	-	-	putative two-component system response regulator
116	-	-	hypothetical protein
118	-	-	D-alanyl-D-alanine carboxypeptidase
120	-	-	transketolase
121	-	-	threonyl-tRNA synthetase
123	-	-	sugar ABC transporter periplasmic protein / family 1 extracellular solute-binding protein
124	-	-	major facilitator family transporter
125	-	-	CMP/dCMP deaminase zinc-binding protein
126	-	-	Mg(2+) transporter mgtE
130	-	-	fructose-1,6-bisphosphatase, class II
132	-	-	hypothetical protein
133	-	-	binding-protein-dependent transport system inner membrane protein
134	-	-	aldo/keto reductase family oxidoreductase
135	-	-	transcriptional regulator
137	-	-	hypothetical protein
144	-	-	ADP-ribosylation/Crystallin J1 / putative membrane protein / ADP-ribosylglycohydrolase
145	-	-	ABC transporter ATP-binding protein
146	-	-	putative nicotinate phosphoribosyltransferase
147	-	-	6-phosphogluconate dehydrogenase

148	-	-	N-acyl-L-amino acid amidohydrolase
149	-	-	NADH:flavin oxidoreductase/NADH oxidase
153	-	-	Alpha or beta hydrolase fold-3 domain protein
155	-	-	ADP-ribosylation/Crystallin J1 / ADP-ribosylglycohydrolase
156	-	-	cell wall endopeptidase
158	-	-	hypothetical protein
159	-	-	5,10-methylenetetrahydrofolate reductase
161	-	-	3-oxoacyl-[acyl-carrier protein] reductase / short-chain dehydrogenase/reductase SDR
162	-	-	5'-nucleotidase sure
163	-	-	binding-protein dependent transport system inner membrane protein
168	-	-	bis(5'-nucleosyl)-tetraphosphatase / metallophosphoesterase
169	-	-	hypothetical protein
171	-	-	DNA-3-methyladenine glycosylase I
172	-	-	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase
175	-	-	hypothetical protein
180	-	-	histidine triad (HIT) protein
181	-	-	MaoC domain-containing protein dehydratase
182	-	-	hypothetical protein
183	-	-	hypothetical protein
185	-	-	thioesterase
186	-	-	cyclase/dehydrase
187	-	-	hypothetical protein

APÊNDICE 4 - CÓDIGO FONTE REFERENTE AO ALGORITIMO BLASTER EM LINGUAGEM JAVA.

```
publicclassBlastExporter {

        publicstaticboolean export(List<HitTextResult>results,String outputFilePath) {

        try {

                File file = newFile(outputFilePath);

                if (file.exists()) {

                        file.delete();

                }

                if (!file.exists()) {

                        file.createNewFile();

                }

FileWriterfw = newFileWriter(file.getAbsoluteFile());

                BufferedWriterbw = newBufferedWriter(fw);

                StringBuildertitle = newStringBuilder();

                title.append("QueryId").append("\t");

                title.append("SubjectIds").append("\t");

                title.append("IdentityPercent").append("\t");

                title.append("PositivesPercent").append("\t");

title.append("AlignmentLength").append("\t");

                title.append("Mismatches").append("\t");

                title.append("GapOpens").append("\t");

                title.append("QueryStart").append("\t");

                title.append("QueryEnd").append("\t");

                title.append("SequenceStart").append("\t");

                title.append("SequenceEnd").append("\t");

                title.append("Evalue").append("\t");

                title.append("BitScore").append("\t");

                title.append("Accession").append("\t");
```

```
                    title.append("Description").append("\t");
                    title.append("Locus").append("\t");
                    title.append("Organism").append("\t");
                    title.append("Source").append("\t");
                    title.append("Version").append("\t");
                    title.append("Keywords").append("\t");
                    title.append("Kingdom").append("\t");
                    title.append("Phylum").append("\t");
                    title.append("Class").append("\t");
                    title.append("Order").append("\t");
                    title.append("Taxonomies").append("\t");
                    bw.write(title.toString());
                    bw.newLine();
                            for (HitTextResultresult : results) {
                            StringBuildersb = newStringBuilder();
                            sb.append(result.getQueryId());
                            sb.append("\t");
                            if (result.getSubjectIds() != null) {
                                    for (String s : result.getSubjectIds()) {
                                            sb.append(s);
if (result.getSubjectIds().size() > 1)
                                                    sb.append(";");
                                    }
                            }
                            sb.append("\t").append(result.getIdentityPercent());
                            sb.append("\t").append(result.getPositivesPercent());
        sb.append("\t").append(result.getAlignmentLength());
                            sb.append("\t").append(result.getMismatches());
                            sb.append("\t").append(result.getGapOpens());
                            sb.append("\t").append(result.getQueryStart());
                            sb.append("\t").append(result.getQueryEnd());
                            sb.append("\t").append(result.getSequenceStart());
```

```
                sb.append("\t").append(result.getSequenceEnd());
                sb.append("\t").append(result.getEvalue());
                sb.append("\t").append(result.getBitScore());
                sb.append("\t").append(result.getAccession());
                sb.append("\t").append(result.getDescription());
                sb.append("\t").append(result.getLocus());
                sb.append("\t").append(result.getOrganism());
                sb.append("\t").append(result.getSource());
                sb.append("\t").append(result.getVersion());
                sb.append("\t");
                if (result.getKeywords() != null) {
                        for (String s : result.getKeywords()) {
                                sb.append(s);
                                if (result.getKeywords().size() > 1)
                                        sb.append(";");
                        }
                }
                sb.append("\t");

                if (result.getTaxonomies() != null) {
                        if (result.getTaxonomies().size() >= 1) {
                                sb.append(result.getTaxonomies().get(0));
                        }
                        sb.append("\t");
                        if (result.getTaxonomies().size() >= 2) {
                                sb.append(result.getTaxonomies().get(1));
}
                        sb.append("\t");
                        if (result.getTaxonomies().size() >= 3) {
                                sb.append(result.getTaxonomies().get(2));
                        }
```

```
                                        sb.append("\t");
                                        if (result.getTaxonomies().size() >= 4) {
                                                sb.append(result.getTaxonomies().get(3));
                                        }
                                        sb.append("\t");
                                        for (String s : result.getTaxonomies()) {
                                                sb.append(s);
                                                if (result.getTaxonomies().size() > 1)
                                                        sb.append(";");
                                        }
                                }
                                bw.write(sb.toString());
                                bw.newLine();
                        }
                        bw.close();
                } catch (Exception e) {
                        e.printStackTrace();
                        returnfalse;
                }
returntrue;
        }
}


publicclassGenBankParser {
protectedstaticfinal String LOCUS_TAG = "LOCUS";
protectedstaticfinal String DEFINITION_TAG = "DEFINITION";
protectedstaticfinal String ACCESSION_TAG = "ACCESSION";
protectedstaticfinal String VERSION_TAG = "VERSION";
protectedstaticfinal String KEYWORDS_TAG = "KEYWORDS";
protectedstaticfinal String SOURCE_TAG = "SOURCE";
protectedstaticfinal String ORGANISM_TAG = "ORGANISM";
```

```
protectedstaticfinal String REFERENCE_TAG = "REFERENCE";

protectedstaticfinal String AUTHORS_TAG = "AUTHORS";

protectedstaticfinal String CONSORTIUM_TAG = "CONSRTM";

protectedstaticfinal String TITLE_TAG = "TITLE";

protectedstaticfinal String JOURNAL_TAG = "JOURNAL";

protectedstaticfinal String PUBMED_TAG = "PUBMED";

protectedstaticfinal String MEDLINE_TAG = "MEDLINE";

protectedstaticfinal String REMARK_TAG = "REMARK";

protectedstaticfinal String COMMENT_TAG = "COMMENT";

protectedstaticfinal String FEATURE_TAG = "FEATURES";

protectedstaticfinal String BASE_COUNT_TAG_FULL = "BASE COUNT";

protectedstaticfinal String BASE_COUNT_TAG = "BASE";

protectedstaticfinal String START_SEQUENCE_TAG = "ORIGIN";

protectedstaticfinal String END_SEQUENCE_TAG = "//";

protectedstaticfinal Pattern headerLine = Pattern.compile("^LOCUS.*");

protectedstaticfinal Pattern sectp = Pattern
        .compile("^(\\s{0,8}(\\S+)\\s{0,7}(.*)|\\s{21}(/\\S+?)=(.*)|\\s{21}(/\\S+))$");

protectedstaticfinal Pattern lp = Pattern                    .compile("^(\\S+)\\s+\\d+\\s+(bp|aa)\\s{1,4}([dms]s-
)?(\\S+)?\\s+(circular|linear)?\\s*(\\S+)?\\s*(\\S+)?$");

protectedstaticfinal Pattern vp = Pattern                    .compile("^(\\S*?)(\\.(\\d+))?(\\s+GI:(\\S+))?$");

protectedstaticfinal Pattern refRange = Pattern.compile("^\\s*(\\d+)\\s+to\\s+(\\d+)$");

protectedstaticfinal Pattern refp =
Pattern.compile("^(\\d+)\\s*(?:(\\((?:bases|residues)\\s+(\\d+\\s+to\\s+\\d+(\\s*;\\s*\\d+\\s+to\\s+\\d+)*)\\))|\\(sites\\))?
");

protectedstaticfinal Pattern dbxp = Pattern.compile("^([^:]+):(\\S+)$");

privateBufferedReaderfileIn;


publicGenBankParser(String path) {

        InputStreaminputStream;

        try {

                inputStream = newFileInputStream(path);

                Reader reader = newInputStreamReader(inputStream, "UTF-8");

                this.fileIn = newBufferedReader(reader);
```

```
            } catch (Exception e) {
                        e.printStackTrace();
            }
}

public List<GenBankResult>getResults() throws Exception {
            List<GenBankResult>results = newArrayList<GenBankResult>();
            GenBankResultr = newGenBankResult();
            while (getSection(r)) {
                        results.add(r);
                        r = newGenBankResult();
            }
            results.add(r);
            returnresults;
}

publicbooleangetSection(GenBankResultresult) throws Exception {
            Listsection = null;
            String sectionKey = null;
do {
                        section = this.readSection(this.fileIn);
                        sectionKey = ((String[]) section.get(0))[0];
                        if (sectionKey == null) {
                                    thrownew Exception("Section key was null");
                        }
                        // process section-by-section
                        if (sectionKey.equals(LOCUS_TAG)) {
                                    String loc = ((String[]) section.get(0))[1];
                                    Matcher m = lp.matcher(loc);
                                    if (m.matches()) {
                                                result.setLocus(m.group(1));
```

```
                }
            } elseif (sectionKey.equals(DEFINITION_TAG)) {
                result.setDescription(((String[]) section.get(0))[1]);
            } elseif (sectionKey.equals(ACCESSION_TAG)) {
                // if multiple accessions, store only first as accession,
                // and store rest in annotation
            String[] accs = ((String[]) section.get(0))[1].split("\\s+");
                String accession = accs[0].trim();
                result.setAccession(accession);
            } elseif (sectionKey.equals(VERSION_TAG)) {
                String ver = ((String[]) section.get(0))[1];
                Matcher m = vp.matcher(ver);
                if (m.matches()) {
                    String verAcc = m.group(1);
                    result.setVersion(verAcc);
                }
            } elseif (sectionKey.equals(KEYWORDS_TAG)) {
                String val = ((String[]) section.get(0))[1];
                if (val.endsWith("."))
                    val = val.substring(0, val.length() - 1);
                val = val.replace('\n', ' '); // remove newline
                List<String>kws = newArrayList<String>();
                for (String kw : val.split(";")) {
                    kws.add(kw.replace('\n', ' '));
}
                result.setKeywords(kws);
            } elseif (sectionKey.equals(SOURCE_TAG)) {
                String source = ((String[]) section.get(0))[1];
                result.setSource(source);
                StringBuildertaxonomySB = newStringBuilder();
                StringBuilderorganismSB = newStringBuilder();
```

```
                        int i = 0;
                        for (String str : ((String[]) section.get(1))[1].split("\n")){
                                if(!str.contains(ORGANISM_TAG) &&str.contains(";")){
                                        taxonomySB.append(str);
                                } else {
                                organismSB.append(str.replace(ORGANISM_TAG, ""));
                                }
                        }

                        String val = taxonomySB.toString();
                        result.setOrganism(organismSB.toString());
                        List<String>taxs = newArrayList<String>();
                        for (String tax : val.split(";")) {
                                taxs.add(tax.replace('\n', ' ').trim());
                        }
                        result.setTaxonomies(taxs);
                }
        } while (!sectionKey.equals(END_SEQUENCE_TAG));

        booleanhasAnotherSequence = true;

        // Allows us to tolerate trailing whitespace without
        // thinking that there is another Sequence to follow
while (true) {

this.fileIn.mark(1);
intc = this.fileIn.read();
if (c == -1) {
hasAnotherSequence = false;
break;
    }
```

```
if (Character.isWhitespace((char) c)) {
continue;
    }
this.fileIn.reset();
break;
        }
// Finish up.
returnhasAnotherSequence;
}

publicbooleancanRead(BufferedInputStreamstream) throwsIOException {
        stream.mark(2000); // some streams may not support this
        BufferedReaderbr = newBufferedReader(newInputStreamReader(stream));
        final String firstLine = br.readLine();
        booleanreadable = firstLine != null&&headerLine.matcher(firstLine).matches();
        stream.reset();
        returnreadable;
}

privateListreadSection(BufferedReaderbr) throws Exception {
        Listsection = newArrayList();
        String line = "";
        String currKey = null;
        StringBuffercurrVal = newStringBuffer();
        booleandone = false;
        intlinecount = 0;
        while (!done) {
                br.mark(320);
                line = br.readLine();
                String firstSecKey = section.isEmpty() ? "" : ((String[]) section.get(0))[0];
                if (line != null&&line.matches("\\p{Space}*")) {
```

```
                        // regular expression \p{Space}* will match line
                        // having only white space characters
                        continue;
                }
                if (line == null|| (!line.startsWith(" ") &&linecount++ > 0 &&
(!firstSecKey.equals(START_SEQUENCE_TAG) || line.startsWith(END_SEQUENCE_TAG)))) {
                        // dump out last part of section
                        section.add(new String[] { currKey, currVal.toString() });
                        br.reset();
                        done = true;
                } else {
                        Matcher m = sectp.matcher(line);
                        if (m.matches()) {
                        // new key
                        if (currKey != null)
                                section.add(new String[] { currKey,currVal.toString() });
                        currKey = m.group(2) == null ? (m.group(4) == null ? m.group(6) : m.group(4)) :
m.group(2);
                        currVal = newStringBuffer();
                        currVal.append((m.group(2) == null ? (m.group(4) == null ? "":m.group(5))          :
m.group(3)).trim());
                        } else {
                                // concatted line or SEQ START/END line?
                                if (line.startsWith(START_SEQUENCE_TAG)||
line.startsWith(END_SEQUENCE_TAG))
                                        currKey = line;
                                else {
                                        currVal.append("\n");
                                        currVal.append(currKey.charAt(0) == '/' ? line.substring(21) :

line.substring(12));
                                }
                        }
                }
```

```
            }
            return section;
}
}

public class GenBankResult {
            protected String locus;
            protected String description;
            protected String accession;
            protected String version;
            protected List<String> keywords;
            protected String source;
            protected String organism;
            protected List<String> taxonomies;

            public String getLocus() {
                        return locus;
            }
            public void setLocus(String locus) {
                        this.locus = locus.replace('\n', ' ');
            }
            public String getDescription() {
                        return description;
            }
            public void setDescription(String description) {
                        this.description = description.replace('\n', ' ');
            }
            public String getAccession() {
                        return accession;
            }
            public void setAccession(String accession) {
                        this.accession = accession.replace('\n', ' ');
```

```
	}
	public String getVersion() {
		returnversion;
	}
	publicvoidsetVersion(String version) {
		this.version = version.replace('\n', ' ');
	}

	public List<String>getKeywords() {
		returnkeywords;
	}
	publicvoidsetKeywords(List<String>keywords) {
		this.keywords = keywords;
	}
	public List<String>getTaxonomies() {
		returntaxonomies;
	}
	publicvoidsetTaxonomies(List<String>taxonomies) {
		this.taxonomies = taxonomies;
	}
	public String getSource() {
		returnsource;
	}
	publicvoidsetSource(String source) {
		this.source = source.replace('\n', ' ');
	}
	public String getOrganism() {
		returnorganism;
	}
	publicvoidsetOrganism(String organism) {
this.organism = organism.replace('\n', ' ');
```

```
        }
}

publicclassHitGenMerger {
        privateGenBankParsergenBankParser;
        privateHitTextParserhitTextParser;
        private List<GenBankResult>genDB;
        private List<HitTextResult>hitDB;
        publicHitGenMerger(String hitFile, String genFile) {
                this.genBankParser = newGenBankParser(genFile);
                this.hitTextParser = newHitTextParser(hitFile);
        }
        public List<HitTextResult>getResult() {
                returnthis.hitDB;
        }
        publicboolean merge() {
                try {
                        this.loadDB();
                        for (HitTextResulth : this.hitDB) {
                                GenBankResultdata = findSubject(h);
                                h.loadData(data);
                        }
                } catch (Exception e) {
                        e.printStackTrace();
                        returnfalse;
                }
                returntrue;
        }
        privatevoidloadDB() throws Exception {
                this.genDB = genBankParser.getResults();
                this.hitDB = hitTextParser.getResults();
```

```
        }


        privateGenBankResultfindSubject(HitTextResultsubject) {
                for (GenBankResultr : this.genDB) {
                        for (String subjectId : subject.getSubjectIds()) {
                                if (subjectId.contains(r.getAccession())) {
                                        returnr;
                                }
                        }
                }

                for (GenBankResultr : this.genDB) {
                        for (String subjectId : subject.getSubjectIds()) {
                                if (subjectId.contains(r.getAccession().replace("_", "|"))) {
                                        returnr;
                                }
                        }
                }
                returnnull;
        }
}









publicclassHitTextParser {
protectedstaticfinalint COLUMN_SIZE = 13;
privateBufferedReaderfileIn;
```

```
publicHitTextParser(String path) {
        InputStreaminputStream;
        try {
                inputStream = newFileInputStream(path);
                Reader reader = newInputStreamReader(inputStream, "UTF-8");
                this.fileIn = newBufferedReader(reader);
        } catch (Exception e) {
                e.printStackTrace();
        }
}

public List<HitTextResult>getResults() throws Exception {
        List<HitTextResult>results = newArrayList<HitTextResult>();
        HitTextResultr = newHitTextResult();
        while (getSection(r)) {
                results.add(r);
                r = newHitTextResult();
        }
        returnresults;
}

publicbooleangetSection(HitTextResultresult) throws Exception {
        try {
        String thisLine;
        while ((thisLine = this.fileIn.readLine()) != null) {
                if (!thisLine.startsWith("#") &&!thisLine.trim().isEmpty()) {
                        String[] r = thisLine.split("\t");
                        if(r.length == COLUMN_SIZE){
                                String queryId = r[0].trim();
                                floatidentityPercent = Float.parseFloat(r[2].trim());
                                floatpositivesPercent = Float.parseFloat(r[3].trim());
```

```
                    intalignmentLength = Integer.parseInt(r[4].trim());
                    intmismatches = Integer.parseInt(r[5].trim());
                    intgapOpens = Integer.parseInt(r[6].trim());
                    intqueryStart = Integer.parseInt(r[7].trim());
                    intqueryEnd = Integer.parseInt(r[8].trim());
                    intsequenceStart = Integer.parseInt(r[9].trim());
                    intsequenceEnd = Integer.parseInt(r[10].trim());
                    doubleevalue = Double.parseDouble(r[11].trim());
                    doublebitScore = Double.parseDouble(r[12].trim());
                    List<String>subjectIds = newArrayList<String>();
                    String[] subjects = r[1].split(";");
                    for(inti = 0 ; i<subjects.length ; i++){
                            subjectIds.add(subjects[i].trim());
                    }
                    result.setQueryId(queryId);
                    result.setSubjectIds(subjectIds);
                    result.setIdentityPercent(identityPercent);
                    result.setPositivesPercent(positivesPercent);
                    result.setAlignmentLength(alignmentLength);
                    result.setMismatches(mismatches);
                    result.setGapOpens(gapOpens);
                    result.setQueryStart(queryStart);
                    result.setQueryEnd(queryEnd);
                    result.setSequenceStart(sequenceStart);
                    result.setSequenceEnd(sequenceEnd);
                    result.setEvalue(evalue);
                    result.setBitScore(bitScore);
                    returntrue;
                }
            }
        }
```

```
} catch (IOExceptione) {
            e.printStackTrace();
}
returnfalse;
}
}



publicclassHitTextResultextendsGenBankResult {
            private String queryId;
            private List<String>subjectIds;
            privatefloatidentityPercent;
            privatefloatpositivesPercent;
            privateintalignmentLength;
            privateintmismatches;
            privateintgapOpens;
            privateintqueryStart;
            privateintqueryEnd;
            privateintsequenceStart;
            privateintsequenceEnd;
            privatedoubleevalue;
            privatedoublebitScore;
            public String getQueryId() {
                        returnqueryId;
            }
            publicvoidsetQueryId(String queryId) {
                        this.queryId = queryId;
            }
            public List<String>getSubjectIds() {
                        returnsubjectIds;
            }
```

```
publicvoidsetSubjectIds(List<String>subjectIds) {
        this.subjectIds = subjectIds;
}
publicfloatgetIdentityPercent() {
        returnidentityPercent;
}
publicvoidsetIdentityPercent(floatidentityPercent) {
        this.identityPercent = identityPercent;
}
publicfloatgetPositivesPercent() {
        returnpositivesPercent;
}
publicvoidsetPositivesPercent(floatpositivesPercent) {
        this.positivesPercent = positivesPercent;
}
publicintgetAlignmentLength() {
        returnalignmentLength;
}
publicvoidsetAlignmentLength(intalignmentLength) {
        this.alignmentLength = alignmentLength;
}
publicintgetMismatches() {
        returnmismatches;
}
publicvoidsetMismatches(intmismatches) {
        this.mismatches = mismatches;
}
publicintgetGapOpens() {
        returngapOpens;
}
publicvoidsetGapOpens(intgapOpens) {
```

```
            this.gapOpens = gapOpens;
        }
        publicintgetQueryStart() {
            returnqueryStart;
        }
        publicvoidsetQueryStart(intqueryStart) {
            this.queryStart = queryStart;
        }
        publicintgetQueryEnd() {
            returnqueryEnd;
        }
        publicvoidsetQueryEnd(intqueryEnd) {
            this.queryEnd = queryEnd;
        }
        publicintgetSequenceStart() {
            returnsequenceStart;
        }
        publicvoidsetSequenceStart(intsequenceStart) {
            this.sequenceStart = sequenceStart;
        }
        publicintgetSequenceEnd() {
            returnsequenceEnd;
        }
        publicvoidsetSequenceEnd(intsequenceEnd) {
            this.sequenceEnd = sequenceEnd;
        }
        publicdoublegetEvalue() {
            returnevalue;
        }
        publicvoidsetEvalue(doubleevalue) {
            this.evalue = evalue;
```

```
        }
        publicdoublegetBitScore() {
                returnbitScore;
        }
        publicvoidsetBitScore(doublebitScore) {
                this.bitScore = bitScore;
        }
        publicvoidloadData(GenBankResultgenBankData) {
                if (genBankData != null) {
                        this.locus = genBankData.getLocus();
                        this.description = genBankData.getDescription();
                        this.accession = genBankData.getAccession();
                        this.version = genBankData.getVersion();
                        this.keywords = genBankData.getKeywords();
                        this.source = genBankData.getSource();
                        this.organism = genBankData.getOrganism();
                        this.taxonomies = genBankData.getTaxonomies();
                }
        }
}
```